WORKING PAPER SERIES*

DEPARTMENT OF ECONOMICS

ALFRED LERNER COLLEGE OF BUSINESS & ECONOMICS

UNIVERSITY OF DELAWARE

WORKING PAPER NO. 2009-02

PRICING CANADIAN AIRPORTS

Joseph I Daniel

*http://lerner.udel.edu/economics/workingpaper.htm

Pricing Canadian Airports

JOSEPH I DANIEL[a1]
[a]Department of Economics, Purnell Hall, University of Delaware, Newark, DE 19716, USA;
e-mail: danielj@lerner.udel.edu; voice: 302-831-1913; fax: 302-831-6968

Congestion pricing of Canada's four largest airports would save between seventy-two and one-hundred-five million dollars annually. Social cost of each aircraft movement would decrease by several hundred dollars at Toronto and Vancouver, and by about fifty dollars at Calgary and Montreal. Toronto currently experiences this congestion in spite of its slot control system. Congestion fees would be less than current weight-based landing fees on average. At projected traffic growth rates, social costs of landings and takeoffs would remain below current levels for at least five years—postponing the need for additional capacity. A stochastic bottleneck model indicates these substantial welfare gains regardless of whether dominant airlines internalize their self-imposed delays. This paper reports equilibrium congestion fee schedules by time of day and calculates equilibrium traffic rates, queuing delays, layover times, and connection times.
Keywords: airport congestion pricing, stochastic queuing, bottleneck model, slot constraints.
(JEL R4, H2, L5, L9)

THE LEADING AIRPORTS in the Canadian National Airport System (NAS) have undergone rapid development and substantial growth in passenger enplanements and aircraft movements since the federal government transferred their operation to local airport authorities during the 1990's. Toronto, Vancouver, Calgary, and Montreal have all substantially modernized their airport terminals, runways, control towers, taxiways, and ground transportation to meet growing demand. Toronto, Vancouver, and Calgary envision construction of additional runways sometime during the next twenty years. All four airports will approach their maximal capacities at their current sites by then.

Toronto's L. B. Pearson International airport is a particularly interesting case, being the largest and most congested airport in the NAS. The airport has among the highest weight-based landing fees in the world ($34.65 per 1000 kg maximum take off weight), and also imposes high Airport Improvement Fees per enplaning ($20) and connecting ($8) passengers. Toronto is the only airport in the NAS currently subject to slot constraints,[2] although the next three largest airports potentially are candidates for slot constraints in the coming decades. As shown below, Toronto experiences significant traffic peaking in spite of its slot constraints, with aircraft imposing around $1000 of external congestion on each other under the current pricing system. Whatever effect Toronto's existing slot reservation system may have, it leaves the airport with substantial amounts of congestion. The results below support the common criticisms that Toronto is overbuilt, overpriced, and inefficient in using its existing capacity. The results cast doubt on the efficacy of administered slot reservation systems as congestion mitigation policies.

This study uses the stochastic bottleneck model of Daniel [10, 11] and Daniel and Harback [12, 13] with data on the airports' scheduled arrival and departure times to calculate equilibrium traffic patterns under weight-based fees and optimal congestion fees. Congestion pricing schedules are calculated under the alternative assumptions that dominant airlines do or do not internalize their self-imposed congestion, because this issue is unresolved in the literature. Daniel and Harback find that dominant airlines do not internalize self-imposed delays at most US airports, while Brueckner [6] and Mayer and Sinai [22] find that they do. These alternative assumptions regarding internalization result in significantly different congestion-fee schedules, but congestion pricing substantially reduces the social costs of delay under either hypothesis. The results also show that making more efficient use of existing capacity by spreading peak demand would postpone the need for constructing additional runway capacity by at least five years. Equilibrium traffic rates, delays, marginal congestion levels, price schedules, and welfare effects of congestion pricing at Canada's four largest airports are reported for both specifications of the model.

In bottleneck models, the purpose of congestion pricing is to reduce delays by smoothing the flow of traffic so that it does not greatly exceed capacity during peak periods. Welfare gains result from more efficient scheduling of existing traffic to reduce delays, not from eliminating some traffic by charging high tolls. In the standard bottleneck model, total demand is fixed while traffic shifts between periods to establish the equilibrium. Most empirical models of congestion pricing, however, specify traffic rates as functions of contemporaneous travel costs, without allowing traffic to shift between periods. The purpose of congestion pricing in such models is to reduce delay by "tolling off" traffic that exceeds socially efficient levels. Canadian airports already impose unusually high charges per enplaned passengers (CEP) in the form of weight-based landing fees and airport improvement fees. Current landing fees alone are higher than optimal congestion fees would be on average. Imposing congestion pricing while rolling back other fees could therefore leave full costs (fees and time costs) of aircraft landings and takeoffs unchanged. While this cost-neutral implementation of congestion pricing would not toll off any more demand, the results in this study show that significant welfare gains would accrue from peak spreading.

The traffic data presented below exhibit rapid periodic peaking associated with hub-and-spoke airline operations. Hub-and-spoke airlines schedule flights in banks of arriving and departing flights that cluster around passenger interchange periods to facilitate connections and reduce layover times. The resulting traffic patterns are significantly different from the gradual morning and afternoon peaking that was common before widespread adoption of hub-and-spoke networks. When airport traffic peaks gradually, the fee schedules also vary gradually over time. When airports have rapid traffic fluctuations associated with hub-and-spoke operations, however, fee schedules must fluctuate rapidly to encourage some aircraft to use off-peak capacity. Most "real world" congestion mitigation proposals are based on models that ignore this rapid peaking of traffic. In these proposals, either variation in congestion fees over time is inadequate, or variation in the number of slots by time periods is inadequate for their prices to vary sufficiently. Whenever price variation is inadequate, queues develop to equilibrate supply and demand.

Unlike standard congestion pricing models, the bottleneck model has endogenously-peaking traffic and explicitly accounts for trade-offs between queuing delay, layover time, and connection time. Traffic rates and queues peak because airlines balance costs of layover delays or missed connections against costs of queuing delays. While some traffic peaking is necessary to optimize connection and layover time, there are two potential sources of excessive peaking that cause inefficiencies. First, non-hub aircraft ignore the delays they impose on the hub airline (and each other) when they schedule aircraft during the hub's flight banks. Second, the hub airline may schedule its own aircraft more closely together than it otherwise would (i.e., ignore self-imposed delays) to exclude non-hub aircraft from operating during its flight banks. This article demonstrates that peak spreading produces significant welfare gains by reducing inefficiencies from either (or both) source(s). To obtain these welfare gains, however, congestion prices must appropriately reflect the dynamic nature of the problem by imposing time-varying fees that have peaking patterns similar to existing traffic rates and queues. Equilibrium congestion fees provide alternative incentives that (largely) replace the role of queuing costs in establishing equilibrium. Optimal congestion fees reduce queuing by spreading out traffic rates so that they do not exceed the service rates.

Slot reservation systems like those used in Toronto and the European Union only attempt to manage demand by hour or half-hour intervals, and consequently lack the precision to prevent congestion delays caused by rapid hub-and-spoke traffic peaking. Toronto is a level 3 coordinated airport that follows the International Air Transport Association's *Worldwide Scheduling Guidelines, 2005, 12th Edition* [19] (WSG) for administration of its slot reservation system. Level 3 airports have demand that exceeds capacity, are unable to restrain demand through voluntarily agreements, and therefore require aircraft to obtain slot reservations to operate. A “slot” authorizes an aircraft to land or takeoff at the airport on a particular day and time, usually designated by an hour or half-hour time window. Slots are typically allocated in series of five or more consecutive days. The fundamental principles of the WSG slot allocation system are to preserve existing usage patterns (subject to a use-it-or-lose-it rule), to limit slot trading to one-for-one exchanges (no sales), to prevent “confiscation” of slots from incumbents for redistribution to new entrants, and to limit slot allocations for new entrants (if at all) to a fraction of new capacity. Twice a year, over 300 airlines worldwide submit Slot Clearance Requests to about 160 airport slot coordinators to substantiate their previous slot usage. After slot coordinators make preliminary slot allocations, over 900 representatives of airports and airlines meet at three-day bi-annual Worldwide Slot Conferences to engage in bilateral one-for-one slot exchange negotiations (see, Gillen, Hendrikssen, et al., [16]). Additional one-for-one exchanges, but not sales, may occur following the conferences. This process must overcome an overwhelming distributional problem that the value of slots depends on combining them with other slots. Airlines must match landing slots with appropriately timed takeoff slots. They must obtain groups of slots to permit banks of arrivals and departures. They must pair slots with appropriately timed slots at the origin or destination airport. They must obtain sequences of slots to support regularly scheduled service with the appropriate frequency. Toronto alone has around 100,000 slots to allocate at each slot conference. Any slot allocation system has to solve a combinatorial optimization problem of enormous dimensionality.

This paper focuses on congestion pricing rather than slot allocations as a solution to airport congestion because the empirical results demonstrate that Toronto's slot system does not regulate aircraft operating times with sufficient precision to eliminate substantial amounts of delay. Slots must have wide time windows to provide them with flexibility, facilitate their exchange, simplify their combination with other slots, and reduce the dimensionality of the scheduling problem. Optimizing delay would require more specific slot windows that grant authority to operate within a moving time window around a particular minute (say fifteen minutes on either side), but this would also greatly increase transaction, administration, and compliance costs. Instead, sets of slots have the same hourly or half-hourly windows to make them more fungible. Consequently, slot constrained airports like Toronto have busy and slack periods within each slot window, resulting in peaking similar to non-slotted airports.[3]

Proponents of slot allocations systems often overlook the complexity of administering slot reservations and the likelihood that they will fail to eliminate large amounts of delay. Slot auctions and slot markets are also administratively complex, relatively inflexible, and subject to high transaction costs. In addition, slot reservation systems may raise entry barriers, facilitate concentration, and increase market power. Congestion pricing, on the other hand, conveys all the necessary information about external costs needed to optimize scheduling while preserving flexibility and freedom of entry. The free-market allocation system in the US ensures open access to airports on a first-come first-served basis. Weight-based landing fees achieve equilibrium of supply and demand using a combination of constant prices and time-varying waiting times. Markets clear based on adjustments of arrival rates and queuing times without the high transaction costs required to administer slot allocation systems. The stochastic bottleneck model used here is essentially a means of replacing waiting time in a queue with marginal cost pricing. Setting the correct price requires determination of marginal social costs based on aircraft operating costs, passenger time values, observed traffic rates, and the dynamics of the queuing system. The queuing model provides analytical expressions for the additional delays imposed on other aircraft for any given traffic pattern. The market can adjust to the resulting time-varying fees just as it currently adjusts to time-varying queuing costs.

Pricing operations at marginal cost and letting the market adjust accordingly does not necessarily require anticipating equilibrium traffic patterns or prices in advance. Price variation would gradually take over the role of queuing time in clearing the market. Slowly phasing in congestion pricing would allow traffic rates to adjust gradually as progressively more of aircrafts' marginal social costs are included in the fees. As congestion fees approach equilibrium levels, queuing time would diminish. In the paper below, computational methods solve the stochastic bottleneck model to demonstrate what equilibrium traffic patterns, queues, and fee schedules should look like at Canadian airports. In practice, however, marginal cost prices can be determined analytically for any given traffic pattern, while letting the market determine demand responses. The costs of additional delays are reasonably straight forward to determine. Aircraft costs and passenger time values are currently determined for a variety of other purposes; for example, in cost benefit analyses to determine optimal airport capacities. Daniel and Harback [12] validates the

stochastic bottleneck model by closely replicating the actual queuing patterns of twenty-seven US airports with a variety of traffic patterns. Calculating the appropriate fees is vastly simpler than administering combinatorial slot auctions or the WSG slot allocation system.

The next section reviews the literature on congestion pricing, internalization of delay, the current slot system, and slot auction and trading systems. Section 3 presents the stochastic bottleneck model while deferring the analytical solutions for marginal social delay times to the Appendix. Section 4 discusses the traffic data and explains the parameterizations of the model. Section 5 illustrates the rapid peaking created by hub-and-spoke airline operations and presents the main results of the paper including the equilibrium traffic, queuing, and pricing schedules for the four airports. It also presents the welfare improvements resulting from congestion pricing and changes in congestion levels given the projected growth in airport traffic over the next twenty years. The paper concludes with a discussion of policy implications based on the results.

The literature

Relatively few articles in the literature on airport congestion pricing attempt to calculate equilibrium fee schedules for actual airports. The literature focuses on theoretical pricing models, pricing of stylized networks, and (most recently) the internalization debate. Carlin and Park [9] estimate external congestion levels for New York's LaGuardia airport circa 1967. Their model assumes that each arrival imposes delays equal to its service time (the time elapsed between consecutive aircraft touchdowns) on every subsequent arrival until the landing queue clears. This approach implicitly assumes a deterministic queuing system similar to those of deterministic bottleneck models. Their model uses data on traffic rates that vary by hour. Koopman [21] develops a stochastic queuing model to calculate delays at New York's J. F. Kennedy airport circa 1970. His model and data also use hourly traffic rates. Carlin and Park and Koopman model traffic patterns as exogenous, so they only estimate the congestion that aircraft impose or experience in the weight-based-fee cases, not the equilibrium congestion-pricing cases.

Borins [5] estimates equilibrium congestion fees at Toronto (Malton) International Airport for actual 1975 traffic and forecast traffic levels into the 1980's. He fits constant elasticity demand functions to determine demand on routes with different stage lengths and times of operation. The model also includes terminal and ground congestion. Full travel price includes ground access time, terminal time, runway time, and whatever fee is assessed. His model has four periods of the day: morning and evening peaks with mid-day and night-time lulls. Cross elasticities of demand allow for shifting between time periods, but not flight distances. His congestion function is based on a (static) steady state queuing model for each of the four periods. Borins estimates congestion fees of about $30 off peak and $150 to $200 during peak periods. His welfare analysis indicates that congestion pricing postpones new capacity requirements by about five years and that Malton is a better airport site than Pickering, the alternative site under consideration at the time. The morning and afternoon peaking pattern with gradual traffic variation is essential to Borins' justification for using static queuing models to determine delays in each period. This approach is reasonable for airport

traffic in the 1960's and 1970's before the proliferation of hub-and-spoke networks, but today's traffic has much more frequent peaking and more rapid fluctuations. Borins [5] is apparently the only previously published academically study that calculates actual congestion prices for Canadian airports.

Morrison [23] and Winston and Morrison [24] calculate airport delays, congestion prices, optimal capacities, and welfare effects from hypothetical congestion pricing at major US airports. They estimate delay times by hour as functions of the hourly number of aircraft landing and taking off, the type of aircraft, the number of runways, and the air-traffic control expenditures at each airport. They model airport demand as a function of hourly congestion prices and delay costs. Welfare gains accrue largely from tolling off aircraft that are unwilling to pay the full social costs of their operations. Their model does not include changes in layover delays or connection times that result when flights shift to less preferred periods.

The airline industry has changed significantly since Borins, Winston, and Morrison's studies in ways that make their models less applicable today. Current data show that hub-and-spoke operations produce traffic patterns that typically fluctuate greatly during any given hour or half hour. Consequently, steady state queuing models and regression models with delays as functions of (half) hourly volume-to-capacity ratios no longer capture airport delay patterns adequately (see, Daniel and Pahwa [14]). Modeling current airport delay patterns requires structural models with state-dependent queuing systems and endogenous traffic rates that adjust continuously (see also, Arnott, et al. [4]).

Vickrey [27] and Arnott, et al. [3, 4] develop bottleneck models in which the timing of traffic adjusts endogenously to minimize the sum of travelers' schedule and congestion delay costs. They apply their models to rush-hour commuting, but the models also apply to hub airports where many aircraft arrive and depart through runway bottlenecks that prevent all aircraft from operating at precisely the beginning or ending of the hub airlines' passenger interchange periods. These standard bottleneck models, however, have deterministic queues that are more appropriate for highway congestion where each vehicle is an infinitesimal part of traffic rather than airport congestion where each vehicle is non-negligible.

Earlier queuing models like Koopman's [21] have simple models of exogenous airport demand, while providing sophisticated probabilistic models of delays. Morrison [23] and Winston and Morrison [24] have credible models of endogenous airport demand, but their delay models are more suited to gradual traffic fluctuations experienced by airports before the proliferation of hub-and-spoke airline networks. This paper applies the model of Daniel [10] that combines stochastic queuing theory (Koopman, [21], Omosigho and Worthington, [25], and Oum, et al., [26]) with a bottleneck model (Vickrey, [27] and Arnott, et al. [3, 4]). The queuing model captures the stochastic nature of arrivals, departures, and queues, that is an essential element of airport capacity missing from deterministic bottleneck models. The bottleneck model provides endogenous adjustment of time-dependent traffic rates. The model is similar to Henderson's [18] highway bottleneck model with flow-congestion in that optimal pricing lengthens peak periods and does not eliminate all congestion, whereas in Vickrey [27] and Arnott, et al. [3, 4] pricing does not change the duration of peak periods and it completely eliminates congestion. The model used here is more applicable to airports than Henderson's model because it has stochastic queuing rather than flow congestion.

This article implements the same model that Daniel and Harback [13] applied to price twenty-seven major US airports. That paper estimated existing delays by regressing flight (or taxi-out) duration of aircraft with different origins (or destinations) on dichotomous variables for minute of arrival (or departure) at the airport. The coefficients of the dichotomous variables represent the component of flight time due to queuing at the runway that varies with the time of day. Daniel and Harback also estimated the values of layover and connection time relative to queuing time using the basic equilibrium condition of the bottleneck model. Their results demonstrated that the stochastic bottleneck model accurately replicates estimated queuing patterns as the no-fee equilibria for a wide variety of airports. Unfortunately, equivalent data on flight duration and operating times in Canada is not available for estimating current queuing patterns or time values. Implementing the bottleneck model, however, only requires data on the number of hub airline flights participating in each passenger exchange and the number and time distribution of non hub aircraft. This study calibrates the model's time values using Daniel and Harback's estimates from Minneapolis-St. Paul airport (MSP). Instead of comparing equilibrium queues against estimated queues, this study validates equilibrium traffic patterns against the scheduled traffic data. While estimated queuing patterns are unavailable, no-fee equilibrium queuing patterns are comparable to those implied by the stochastic queuing model applied to the actual scheduled traffic rates.

This article compares bottleneck equilibria when dominant airlines either internalize or ignore their self-imposed delays. Daniel [10] notes that dominant airlines might internalize their self-imposed delays. He performs a series of empirical tests using tower log data from MSP airport that largely rejects the internalization hypothesis. In spite of this, many subsequent researchers find the theoretical argument that dominant airlines should internalize their self-imposed delays to be more convincing. Brueckner [6] and Mayer and Sinai [22] both find weak but statistically significant evidence that airport delays decrease as airport concentration increases. They argue that this relationship is due to internalization by dominant airlines and that congestion fees should reflect internalization of self-imposed delays. Brueckner proposes that dominant airline fees should be inversely proportional to their share of aircraft operations at the airport. Daniel and Harback [12] conduct specification tests similar to Daniel [10], for twenty-seven major hub airports and find that dominant airlines do not internalize self-imposed delays at most major U.S. airports, but they may internalize delays at some airports. Brueckner and Van Dender [6] shows that conventional congestion models can generate internalizing or non-internalizing behavior, depending on whether dominant airlines follow Nash or Stackelberg strategies. This article remains neutral about the internalization issue by calculating congestion fees and welfare effects for both specifications. The results illustrate how traffic patterns and price schedules differ depending on these behavioral assumptions.

This paper primarily focuses on pricing to alleviate congestion. An interesting empirical result, however, is that Toronto's slot reservation system fails to eliminate a substantial amount of congestion. Toronto, Europe, and the US currently implement administered slot allocations rather than market-based slot allocations. Gillen, Henriksson, et al. [16] recommends that slots at Toronto (and any NAS airports subject to future slot controls) be auctioned in bundles with slots expiring after a fixed period. Slots would

be resalable. *Jones, Holder, et al. [20] evaluates similar proposals for Europe in their study* assessing the effects of different slot allocation schemes for the European Union. Whalen, Carlton, et al., [32] proposes that JFK, LaGuardia, and Newark airports annually auction ten percent of their slots, with slots subject to expiration and re-auctioning every ten years. Slots would be freely marketable between auctions. Brueckner and Van Dender [7] and Verhoef [31] develop models in which such slot auctions with subsequent marketability achieve optimal allocations of slots. Verhoef's results limit optimality to situations in which airlines do not have market power. Bruekner's model explicitly assumes airlines have no market power.

Auctions of marketable slots would undoubtedly improve slot allocation systems relative to the WSG approach, but they do not address the conflict between long slot windows necessary for marketability and precise slot windows necessary for optimal peak spreading. Whalen, Carlton, et al., [32] explicitly mentions the problem of demand peaking within the slot periods due to imprecise specification of operating times, but recommends half-hour slot windows to facilitate auctions and sales by reducing the number of potential slot bundles. They suggest revisiting the possibility quarter-hour slot windows if peaking persists. Having continuously varying congestion prices with open access to airports avoids this problem. Airlines already solve the problem of scheduling their operations at large numbers of airports subject to rapidly varying queuing times, but these queuing times provide the wrong signals for social optimality because they do not include congestion externalities. Congestion pricing replaces continuously varying queuing costs with continuously varying prices. Whalen, Carlton, et al., [32] dismisses congestion pricing because they argue market clearing prices are difficult to determine, but *Jones, Holder, et al. [20] details the enormous complexities of* combinatorial auctions that Whalen, Carlton, et al. favor. The remainder of this paper demonstrates how to calculate time varying marginal social costs of aircraft operations at Canadian airports. The concluding section discusses how to implement congestion pricing by dynamically adjusting prices in response to observed changes in demand without having to anticipate equilibrium traffic patterns or prices.

The model

This paper applies the stochastic bottleneck model described in Daniel and Harback [12] to calculate equilibria with weight-based fees and congestion prices at Canada's four leading airports. This section describes the model, while Appendix A presents its mathematical specification.

The standard deterministic bottleneck model describes the endogenous timing of trips in response to delays or fees as vehicles flow through bottlenecks with limited capacities that prevent travelers from all arriving at their destinations at precisely their preferred times. Some travelers must complete their trips early and others must complete their trips late. A deterministic queue develops at the bottleneck whenever the traffic rate exceeds the bottleneck capacity. Travelers have time values associated with queuing time, early trip-completion time, and late trip-completion time. By assumption (and without loss of generality), the time required to traverse the bottleneck is the only time needed to travel between origins and destinations. The equilibrium requires that arrival rates at the bottleneck adjust endogenously so that the

rates of change in queuing-, early-, and late-time costs sum to zero throughout the busy period, for identical travelers. Increasing queuing-time costs just offset decreasing early-time costs for early travelers, and decreasing queuing-time costs just offset increasing late-time costs for late travelers. The standard application of this model is to the morning commute to work, but the model is also applicable to hub airport traffic. Airlines want their aircraft to exchange passengers quickly and reliably with one another at their hub airports. Runway capacity prevents aircraft from all landing or taking off at precisely the same time. Some aircraft arrive early with little queuing but have long layovers, some arrive close to the exchange time have shorter layovers but longer queuing delays, and some arrive late as queues are decreasing but risk missing connections. Airlines adjust their schedules so that in equilibrium no aircraft can reduce the sum of its layover, queuing, and connection costs by changing its intended time of operation. Aircraft with the same time values and operating-time preferences must have the same total cost in equilibrium.

Several modifications to the standard model are necessary to realistically describe hub-and-spoke airport traffic. First, there are significant random shocks to aircraft operating times. Travel time is affected, for example, by headwinds and tailwinds, other weather conditions, and directness or circuity of flight paths. Aircraft schedulers should use the probability distribution on random shocks to operating times to calculate expected layover and connection time costs, rather than assuming aircraft always operate at their intended times as in the standard deterministic model. The distribution of these random shocks is modeled using observations from Daniel and Harback [13] on differences between actual aircraft-operating times and mean operating times per flight number over the sample period at MSP airport.

Random aircraft operating times also affect the queuing system, particularly when each aircraft is a non-negligible unit of traffic (unlike automobiles). Stochastic queuing theory treats traffic rates and queue lengths probabilistically. Arrivals at the queue follow Poisson distributions with time-varying expected traffic rates. Discrete probability distributions on queue lengths describe the states of the queuing systems. The rate of change in the state of the queuing system depends on both the current traffic rate and the current state of the queue. Daniel and Pahwa [14] show that stochastic queuing systems are more successful than deterministic queuing systems at modeling the evolution of airport queues over time. The specific queuing model used here is classified as *M(t)/d/s/K*; indicating that it has time-dependent Poisson traffic rates, a deterministic service rate, a maximum queue size, and multiple servers (runways). This specification is chosen for its realism and computational simplicity.

Daniel [10] combined the bottleneck model with a stochastic queuing system to endogenize traffic rates. The stochastic bottleneck model has the same basic structure as the deterministic version, but its equilibrium requires that traffic rates adjust until the *expected* rates of change in layover, connection, and queuing time costs sum to zero throughout the busy period. In equilibrium, there is no further incentive for airlines to change the intended operating times of any aircraft. In the model, as in reality, airlines schedule aircraft without knowing the random shocks their aircraft will experience or the precise state of the queue at the times of service. The airline calculates expected layover, connection, and queuing times by using the distribution of random shocks and the distribution of queue lengths. Deterministic and stochastic queuing

models can behave very differently. Stochastic queues have positive expected lengths whenever the traffic rate is positive, whereas deterministic queues do not develop as long as the traffic rate is at or below the service rate. In the priced deterministic equilibrium, therefore, traffic rates exactly equal the service rate throughout the entire busy period, with no queue developing. The length of the busy period is constant. In the stochastic bottleneck model, however, airlines always trade off layover, connection, and queuing times. Some queuing delay is unavoidable. Mayer and Sinai [22] refer to this unavoidable delay by saying, "… not all airport delays are evil." Congestion pricing in the stochastic bottleneck model distinguishes between good (efficient) queuing delays that are required to optimize layover and connection times and evil (inefficient) queuing delays caused by externalities. Pricing only seeks to eliminate the evil external delays.

Congestion pricing in the stochastic bottleneck model eliminates inefficiencies from purely external delays by spreading out traffic and reducing peak traffic rates. The Nash-dominant airline specification assumes that dominant airlines already internalize the additional delays their aircraft impose on each other. They schedule aircraft to minimize their combined costs of any fees and layover, connection, and queuing times, taking other aircraft schedules as given. Their first-order necessary conditions consist of an equation for each of their aircraft that includes the marginal costs it experiences directly plus its indirect effect on marginal costs of its other aircraft. The airport authority sets Nash-dominant airline fees equal to the additional delays that they impose on other airlines' aircraft. Non-dominant aircraft fees equal the additional delays that they impose on one another and on dominant aircraft. In this specification, welfare gains result primarily from shifting non-dominant aircraft out of the dominant airlines' peak-traffic periods.

The Stackelberg-dominant airline specification, on the other hand, ignores self-imposed congestion in unpriced equilibria because airlines realize that the aggregate traffic patterns and congestion levels are the same in any bottleneck equilibrium regardless of when dominant airlines schedule their own aircraft. Any attempt by dominant airlines to internalize delays by spreading out their traffic simply results in trading places with non-dominant aircraft. Stackelberg dominant airlines schedule their aircraft to minimize their individual aircraft's own costs of any fees and layover, connection, and queuing times, subject to the best responses of non dominant aircraft. Their first-order necessary conditions consist of an equation for each aircraft that includes the marginal costs it experiences directly but assumes it has no marginal effect on its airline's other aircraft because they experience the same delay in any Stackelberg equilibrium. The airport authority sets Stackelberg-dominant airline fees equal to the additional delays their aircraft impose on other airlines' aircraft *and* on one another. The non-dominant aircraft fees are the same as before.

If the same layover and connection time values were applied to both specifications, then the specifications' unpriced equilibrium traffic patterns would be different because of different internalization behavior, while the specifications' priced-equilibrium traffic patterns would be identical because they fully internalize delay in both cases. The observed (unpriced) traffic patterns, however, are (necessarily) the same, so the specifications *must have different layover and connection time values* to approximate this observed traffic pattern. It follows that the priced traffic patterns of the two specifications will be different.

In both specifications, pricing produces welfare gains from shifting non-dominant traffic out of the dominant airlines' flight banks. The non-internalizing specification, additionally, has excessive delays that dominant aircraft impose on each other in the unpriced equilibrium. Congestion prices reduce these self-imposed delays by enabling dominant airlines to optimally spread out their traffic while discouraging non-dominant aircraft from shifting back into the peak periods. While the Nash specification treats dominant airlines' self-imposed delays as part of the optimal tradeoff between layover, connection, and queuing times, the Stackelberg specification treats these delays as though they are external and should be priced. The Stackelberg specification has more external delay but the Nash specification has higher values of layover and connection times. It is indeterminate which specification generates higher welfare gains from congestion pricing. This explains how it is possible to obtain the empirical result that the internalizing specification generates higher welfare gains than the non-internalizing result.

While the model provides closed-form, analytical expressions for the probability distributions on queues and expected marginal delays, it requires computational methods to find the equilibrium traffic rates. For any initial schedule of operations, the model can calculate the full cost (layover-, connection-, and queuing-time costs plus whatever fee is imposed) of scheduling aircraft of a particular type during any service interval. Different aircraft types allow for different values of layover, connection, and queuing times and different distributions of preferred arrival times. Dominant hub airlines and their code affiliates have a single preferred operating time for each arrival or departure bank that is the start or end of the passenger interchange period. All non-dominant aircraft have uniformly distributed operating-time preferences within each bank. The computational algorithm iteratively reschedules dominant aircraft from periods with above average and increasing expected costs to periods with below average and decreasing expected costs until all aircraft of a given type have the same (minimum) cost. Depending on the specification, these expected costs exclude or include the additional delays imposed on other dominant aircraft. The algorithm iteratively reschedules non-dominant aircraft until the first differences of their expected costs sum to zero for all aircraft. Aircraft categories include: major dominant, regional (code sharing) dominant, major non-dominant, regional non-dominant, and other (mostly unscheduled general aviation). Each category has its own time values that account for different aircraft sizes.

Although the stochastic model is more complex computationally than the deterministic model, the models have essentially the same structure with the substitution of *expected* layover, connection, and queuing times for their corresponding deterministic values. The stochastic queuing system takes two additional parameters: namely the number of servers (runways) and a maximal queue size. Airport data determines the number of runways. The maximal queue size is necessary purely for computational purposes and is large enough that there is infinitesimal probability of reaching it. Both the deterministic and stochastic models require parameters for the length of service intervals. Traffic rates are endogenous in both cases; the deterministic system has only two traffic rates, early and late, while the stochastic queuing system has continuously varying traffic rates. Consequently, the stochastic model has no more free

parameters than the deterministic model. The additional complexity of the stochastic model is confined to its probabilistic queuing system.

The Data

The data on scheduled and actual operating times comes from airline web sites that report all scheduled aircraft movements at the four airports on April 1, 2008. They report scheduled and actual times of operation, the origin or destination airports, and the arrival or departure gates. Transport Canada's *Aircraft Movement Statistics*, *Annual Report 2005* [29] and the *2007 Annual Reports* of the respective airports provide data on unscheduled operations. *Aircraft Movement Statistics* does not report the times of operation, but it tabulates aircraft by types, weights, and operators. By assumption, unscheduled aircraft are distributed proportionately to the size of each flight bank (including dominant and non dominant aircraft). Their preferred operating times are uniformly distributed within each bank. Non dominant scheduled aircraft are also assumed to have uniformly-distributed preferred operating times within bank periods. Dominant aircraft are assumed to have a common most preferred operating time at the start or end of passenger interchange period. The bottleneck model determines the start or end of the interchange as the time when the aircraft experiencing the longest queue completes service.

This is the best data that is publicly available, but it is inadequate to statistically estimate actual queuing patterns or the values of layover and connection time relative to queuing time. Daniel and Harback [13], however, estimates time values for a wide variety of US airports and finds that the values are reasonably consistent for similar types of airports. Time values for MSP airport provide the basis for those used here, because MSP is reasonably similar in size and location to the Canadian airports. These time values determine the rates (slopes) of increase and decrease of the traffic peaks, so comparing equilibrium traffic patterns with scheduled traffic peaks provides an alternative validation of the time values and the model's ability to replicate actual traffic patterns. Estimates of actual queuing patterns throughout the day would also be useful for calibrating service rates to more precisely replicate actual queuing patterns. This study uses service rates from the airports' *Master Plans* [1, 8, 17, 30] that report the maximal number of aircraft movements they can perform per hour under various conditions. Airports use these service rates to determine future capacity requirements. Daniel and Harback's work with US data verifies that these are good estimates of the queuing systems' service rate parameters. The stochastic queuing model accurately replicates Daniel and Harback's estimates of actual queues. Service rate elasticities are reported below to indicate the sensitivity of the results with respect to service rates. Aside from problems estimating the time values and actual queuing times, the schedule data is entirely adequate to determine the equilibrium traffic patterns, queues, and congestion fees.

Economic Values for FAA Investment and Regulatory Decisions, a Guide, Table 3-4, p 3-7 [15] (aka, *Critical Values*) provides aircraft operating-cost data to quantify Daniel and Harback's estimates of layover- and connection-time values relative to queuing-time values. *Critical Values* specifies the methodology and economic values to be used in investment and regulatory decisions of the FAA. Similar

documents specify cost-benefit methodologies for Canada, but do not provide precise values (see, *Benefit-Cost Analysis Guide*, Treasury Board of Canada Secretariat, [28]). *Critical Values* compiles the operating costs of aircraft by seating capacity, body type (narrow or wide), and number and type of engines as reported by the airlines on FAA Form 41. The data include numbers of aircraft of each type, passenger capacity, load factor, crew size, and fuel consumption. Different variable costs of aircraft operation account for different rates of fuel consumption while queuing in the air versus on the ground. The passenger time value is $28.50 per passenger hour. The average aircraft and crew have values per block hour of $1245 on the ground and $2096 in the air. Average passenger capacity is 157 and average load factor is 72% of capacity, giving a passenger value per block hour of $3233. Since these values vary nearly proportionately with aircraft size, they can be scaled to represent the queuing-time value of the average aircraft for each airline group at each airport as given in Transport Canada's *Aircraft Movement Statistics, 2005* [29]. Finally, values of layover and connection times are the aircraft queuing-time values multiplied by the layover- and connection-time regression coefficients from Daniel and Harback [12] for MSP airport.[4] The exchange rate between US and Canadian dollars at the time of the traffic observations, April 1, 2008, was very close to one, so no exchange-rate conversion is necessary.

The Results

Figures 1 and 2 compare the model's equilibrium arrival rates with the actual scheduled arrival times at Toronto, Vancouver, Calgary, and Montreal for the non-internalizing and internalizing specifications.[5] The corresponding information for departures is shown in Appendix B, Figures B.1 and B.2. The observed traffic patterns, shown in lighter gray, illustrate the rapid fluctuations associated with operation of hub-and-spoke networks. Precise data on scheduling is clearly necessary to describe the extent of this peaking. Aggregation by hour or half-hour would eliminate most traffic fluctuation. Traffic at Toronto peaks about every half hour, with traffic rates varying from nearly zero to well above the service rate during almost every bank. Existing slot constraints obviously do not eliminate this peaking, but the constraints may cause more and smaller flight banks by limiting the number of operations within each slot window. If slot constraints are binding on the number of aircraft in the banks, then dominant airlines must prefer larger banks with lower layover costs but more queuing time. Slot constraints may actually increase social costs by preventing efficiently sized flight banks. Vancouver is almost as congested as Toronto per aircraft, but has two-thirds as much traffic. It does not have slot constraints and has about half as many banks. Its traffic pattern is also typical of hub-and-spoke operations, with rapid traffic fluctuations ranging from zero to above capacity within half-hour intervals. Similar observations apply to Calgary and Montreal, but they both have less than half Toronto's traffic. Calgary has two dominant airlines and therefore more banks than Vancouver or Montreal.

The actual traffic data is highly consistent with the bottleneck model. The modeled traffic peaks, plotted in black, all have similar structure and timing as the actual scheduled traffic patterns. The shapes of traffic peaks depend on costs of layover and connection time relative to queuing time, the distribution of

preferred operating times, the number of aircraft in a bank, and the service rates. The time values used here, from MSP, are typical of hub airports and produce traffic peaks similar to actual peaks at the Canadian Airports. Minor differences between the actual and modeled traffic peaks result in part from airlines reporting their scheduled operating times rounded to five or ten minute values, introducing some spikes and pits in the schedule data. Traffic rates would be smoother for data on rates of arrival at the queues averaged over several days, but this data is unavailable in Canada. The timing of the modeled traffic peaks depends on the time of dominant aircrafts' most preferred operating time in each bank. The bottleneck equilibrium requires that the aircraft with the longest queuing delay must complete service exactly at its most preferred time. These times are calculated for each bank as the time of the maximal queue plus the waiting time required to complete service. Preferred operating times for non hub aircraft are uniformly distributed over the bank. The number of aircraft in each bank is simply the number of operations preformed between the valleys (minimum traffic rates) on either side of the peak. The service rates determine the heights at which traffic rates plateau. Roughly speaking, unpriced traffic rates tend to level off close to the service rates.

The model fits Toronto and Vancouver's traffic particularly well, as their traffic has the clearest hub-and-spoke bank structure. Toronto has many more banks than the other airports, possibly due to slot constraints, but the structure of the peaks is still highly consistent with the model. Toronto may also have more banks because it has more international traffic than any other Canadian airport and separates aircraft into sets of banks serving domestic, European, and US markets. Toronto's traffic clearly illustrates that slot constraints as currently administered do not eliminate demand peaking. Calgary has the second most banks due to having two dominant airlines, Air Canada and WestJet. Calgary's banks are less regularly spaced and crowd against each other. Daniel and Harback [13] found similar patterns at US airports with multiple dominant carriers such as Boston, Chicago, Dallas-Fort Worth, Newark, and others. Montreal has the weakest hub among the four airports. Its traffic peaks are the least distinct, but still show regular peaks and valleys.

The model's unpriced bottleneck equilibria that use the correct specification of internalizing behavior should match the scheduled traffic patterns best. The two specifications use different estimates of time values because the rate at which aircraft trade off layover and connection time for queuing time depends on whether queuing time includes indirect (imposed on other aircraft) delays in addition to direct (experienced by the aircraft) delays. Precise statistical tests to determine the correct specification require data on actual operating times that are unavailable for Canada, but such tests for US airports generally favor the atomistic specification. Casual comparison of traffic patterns at each of the Canadian airports does not appear to clearly favor either specification.

Figures 3 to 6 illustrate the non-internalizing daily traffic patterns, queue lengths, and congestion levels with and without congestion pricing for each of the airports. Figures 7 to 10 illustrate the same information for the internalizing case. The top panels of each figure show the modeled traffic patterns before and after imposition of the congestion prices. Traffic rates flatten out under congestion pricing, spreading peak rates into periods between peaks. Peak spreading is more evident for the non-internalizing

cases because they have lower layover and connection time values relative to queuing time. The internalizing cases require higher layover and delay values to create the same degree of peaking when airlines internalize their self-imposed delays. For both specifications, unpriced traffic rates peak at about the service rate, causing stochastic queues to grow rapidly. The service rates for Toronto, Vancouver, Calgary, and Montreal are 1.25, 0.73, 0.59, and 0.50 aircraft per minute. By internalizing the high costs of extra delay that aircraft operating near the peak impose on others, congestion pricing shifts operations away from the peaks. Relatively small reductions in peak traffic can significantly reduce delays because expected queues increase at increasing rates as traffic rates approach the service rates.

The second panel of each figure illustrates the change in expected queue lengths resulting from congestion pricing. In the non-internalizing cases, peak queue levels under congestion pricing are roughly half the unpriced levels, but they decrease less rapidly at the edges of the peaks. In the internalizing cases, the reduction in queuing is not as dramatic but peak queuing levels are clearly reduced. None of these airports is highly congested at this time, but the qualitative characteristics of the queues remain similar as daily traffic volumes and congestion increase. Note that queuing is not completely eliminated by pricing in the stochastic bottleneck equilibrium because any positive traffic rate necessarily causes some expected queuing. This is different from the deterministic bottleneck model where traffic rates at or below the service rates cause no congestion. The internalizing specification has less opportunity to reduce queues because there is less uninternalized delay in the unpriced equilibrium. The changes in queuing delay times are quantified below.

In the bottom panel of each figure, the unpriced congestion levels indicate the marginal delay that an operation imposes on other aircraft under the current weight-based pricing system. The priced congestion levels are the optimal equilibrium congestion fee schedules as functions of the time of operation. For the non-internalizing specification, there is a common fee structure that should be imposed both dominant and non dominant aircraft in the priced equilibria. For the internalizing specification, the graphs depict the ideal (internalizing) fee structures for each dominant and non-dominant aircraft type. For the non-internalizing case, the unpriced external congestion levels are typically many times higher than the equilibrium congestion fees because traffic has not adjusted to internalize congestion. Treating traffic rates endogenously demonstrates how dramatically congestion pricing decreases congestion levels in the non-internalizing case. For the internalizing specification, the difference between unpriced and priced external congestion is much smaller. Dominant airlines already internalize their self-imposed congestion, so there is less external congestion for pricing to internalize. Assuming internalization of self-imposed delays by dominant airlines, the ideal congestion prices charge non-dominant aircraft for congestion imposed on all other aircraft, while they charge dominant airlines only for congestion imposed on other airlines' aircraft. Depending on market shares, dominant airline fees can be much lower than non-dominant fees.

The airports' fee structures are qualitatively similar. They increase and decrease almost linearly over time, and reach their peaks just before the beginning of each interchange period. The fees fall to zero between peaks provided that the interchange periods are sufficiently separated from one another

(Vancouver, Calgary, and Montreal) otherwise the queues carryover a base level from one bank to the next that trends up and down with the diurnal variation in demand (Toronto). The dominant airlines' ideal fees in the internalizing case are lower than the non-internalizing fees and appear out of phase with their traffic peaks. These fees are lowest in the center of the banks and relatively higher at the edges because external congestion is highest where non-dominant aircraft operate. Calgary's fee schedule is somewhat more erratic than the others because it has two airlines (Air Canada and Westjet) with separate bank schedules.

The highest congestion fees are higher than current weight-based landing fees, but as shown below, the average full costs (including time costs) of aircraft operations under congestion pricing are generally less than those under weight-based fees. Higher peak fees offset lower layover or connection costs, leading to constant costs over the busy periods. The fees must vary over time to reduce inefficient peaking of the queues. Many "congestion pricing" proposals attempt to "simplify" the fee structures by assigning flat fee surcharges over the morning and evening peak periods. Flat fees, while simple, do not address the causes of inefficient demand peaking and queuing. Toronto illustrates that slot constraints have a similar problem. The slot windows are evidently not precise enough (or inadequately enforced) to prevent substantial peaking of the traffic within the windows. Slot constraints may succeed in producing smaller, more frequent traffic peaks, but they cannot prevent the traffic from peaking with the same pattern as unpriced bottleneck equilibrium. The fee schedules proposed here vary nearly linearly over time as functions of when aircraft join the landing or takeoff queues. These linear fee schedules are simple enough for the airlines to comprehend and are not unduly costly for airports to administer.

Tables 1 and 2 report the quantitative effects of congestion pricing for the non-internalizing and internalizing specifications at the four airports. The tables categorize aircraft by type, including; dominant airlines and their code-affiliated aircraft, all other scheduled aircraft, unscheduled commercial aircraft, and unscheduled private aircraft (if any). The first column of Table 1 shows the number of aircraft that land at the airport during the day. Toronto, Vancouver, Calgary, and Montreal are the largest Canadian airports by number of aircraft movements (landings and takeoffs) in 2008. Winnipeg International is the next largest airport in the National Airport System with about sixty percent as much traffic as Montreal. This paper focuses on the four largest airports in the NAS because they are the only ones with appreciable congestion levels. Toronto has by far the most aircraft movements with 656 pairs of landings and takeoffs. Vancouver has 446, Calgary has 314, and Montreal has 285. The number of operations varies seasonally with about 20% more in summer and 20% less in winter. Toronto has successfully shifted nearly all noncommercial traffic to reliever airports.

The second column of Table 1 shows the approximate weight-based landing fee on average for each airline group based on the distribution of aircraft weights reported in Transport Canada's *Aircraft Movement Statistics, 2005* [29]. Canadian airport revenues come primarily from landing fees, general terminal fees, airport improvement fees, parking fees, and concessions and rentals. Toronto's landing fee is among the highest in the world[6] ($34.65/tonne MTOW)[7] and is much higher than those of Vancouver ($3.24 to 4.86)[8], Calgary ($4.14 to $7.38)[9], and Montreal ($13, estimated).[10] Toronto also has airport

improvement fees of $20 per originating passenger and $8 per connecting passenger. Vancouver's airport improvement fee is $15 per originating passenger and Calgary and Montreal both charge $20 per originating passenger.[11] Air Canada and WestJet have higher proportions of domestic and regional aircraft than other scheduled service, resulting in lower weight-based fees on average. The size of Toronto's average weight-based fee, about $3600, is controversial and arguably due to excessive infrastructure and/or high land rents charged by the federal government. Montreal's average fee, $1000, is also high relative to other North American airports, while Vancouver and Calgary charge about $400 on average, which is closer to normal. Unlike US airports, Canadian airports are not subsidized by the Federal Government.

The third column of Table 1 shows the average congestion fees for each airline group. Average congestion fees are generally lower than current weight-based fees, except for those of some unscheduled aircraft. The average congestion fee for Air Canada and other scheduled aircraft at Toronto is just under $900, while fees for unscheduled aircraft average about $500 because they shift further to the edges of the traffic peaks. At Vancouver, scheduled aircraft fees average about $460, while unscheduled aircraft pay about $180 on average. Scheduled aircraft at Calgary and Montreal pay around $200 and unscheduled aircraft about half that much. Comparing these monetary costs of weight-based and congestion fees ignores changes in layover, connection, and queuing time costs.

The full costs of fees and time appear in columns four and five. Imposition of congestion pricing has a large effect at Toronto because it has such high weight-based fees. For scheduled aircraft, full costs fall by over $3000 on average, while unscheduled aircraft pay about $100 more. The effect on full costs for scheduled aircraft is smaller at Vancouver, decreasing about $200, while the change in unscheduled aircraft costs is about the same. Most aircraft at Calgary and Montreal experience several hundred dollar decreases in full costs. Own-price elasticities of demand with respect to landing and takeoff costs should be small because these costs are typically less than five percent of total flight costs. Cross-price elasticities of demand are very difficult to estimate, but are unlikely to be significantly greater than the own-price elasticities. It is reasonable to treat demand for airports as fixed, because the elasticities are small and airports can compensate for changes in full cost by adjusting other airport fees, such as terminal fees and airport improvement fees. Congestion pricing can be implemented in a cost-neutral way.

Columns six and seven show expected queuing times under weight-based pricing and congestion pricing. Again, this comparison includes only part of the change in delay by excluding layover and connection time, but queuing delays are particularly important because they are the most unnecessarily expensive. At Toronto and Vancouver, average queuing delays for scheduled aircraft fall from nearly four minutes to just over two. Average queuing times at Calgary and Montreal decrease by one to one-half minutes. None of the airports is highly congested at this time, so reductions in average queues are modest. Peak queuing levels at the airports, however, would decline from more than five minutes to close to two except at Calgary where peak queues are currently only about 3.5 minutes. As traffic increases over time, congestion increases rapidly and all four airports face significant congestion within the next twenty years. Traffic growth and capacity are addressed later in this section.

The remaining columns of Table 1 summarize the welfare gains per aircraft, airline group, and airport. Column eight shows the change in direct time costs resulting from congestion pricing, net of the changes in fees. There is no unique way to apportion the social savings among aircraft, but this column illustrates that scheduled aircraft at all four airports have lower time costs, while the unscheduled aircraft have increased time costs because they shift further from their preferred operating times. Changes in fees are zero-sum transfers between airlines and airports, so there is a net social gain resulting from the time savings. Dead weight loss from delay under weight-based pricing is converted into airport revenue under congestion pricing. Congestion pricing recovers about $180 of delays per aircraft at Toronto, and $113 at Vancouver. Calgary and Montreal recover about $50 per aircraft. Columns nine and ten show the effects of completely replacing weight-based fees with congestion fees. Usually the airlines' costs go down more than airports' revenues, so the airports could make the switch to congestion pricing Pareto improving by retaining some of the revenues from weight-based fees. The final column show the daily welfare gains from congestion pricing of roughly $120,000 for Toronto, $50,000 for Vancouver, and $15,000 at Calgary and Montreal. These savings amount to twenty to thirty percent of the congestion fee revenues.

Table 2 presents results analogous to Table 1 for the alternative specification in which dominant airlines do internalize their self-imposed delays. In principle, either specification can result in larger welfare gains from congestion pricing, even though the unpriced internalizing specification has less uninternalized delay. Queuing time values depend on aircraft operating costs that are the same across specifications, but layover delay and connection time values must be higher in the internalizing specification to generate the same queuing levels when dominant airlines internalize. One or the other specification generates incorrect estimates of time values and equilibrium traffic patterns. Which specification is correct may vary by airport. Both the unpriced baselines and priced equilibria traffic patterns are wrong for the incorrect specification. When comparing the dollar values across the two specifications in Tables 1 and 2, keep in mind that they do not use the same time values.

Columns one and two of Table 2 show the weight-based and congestion fees for the specification with internalizing dominant airlines. In this specification, the primary role of congestion pricing is to price non-dominant aircraft out of the dominant airlines' flight banks that interchange passengers at hub airports. Congestion fees are typically lower for dominant airlines than for other major airlines. Unscheduled aircraft, however, are smaller on average and shift further toward the edges of the traffic peaks so they may impose less external delays than commercial airlines. Average congestion fees at Toronto are $481 for dominant aircraft and $1,022 for others; and at Vancouver, $385 for dominant aircraft and $734 for others. At Calgary and Montreal, dominant aircraft pay about $150 and other aircraft pay about $250 on average.

Columns three and four of Table 2 show the changes in time costs and fees per aircraft for the internalizing specification. The time costs exclude the indirect costs that dominant aircraft impose on each other because these costs are counted as direct cost of other dominant aircraft. At Toronto, the full landing and takeoff costs decrease from $5,524 to $2,273 for dominant aircraft, and from $6,313 to $2,806 for others; at Vancouver full costs decrease from $2,129 to $1643 for dominant aircraft, and increase from

$1,443 to $1,591 for others.; and at Calgary and Montreal full costs change from about $1,600 to $1,400 for dominant aircraft, and from $1,500 to $2,100 for others. An important policy concern is that the internalizing specification often leads to lower fees and higher benefits for dominant airlines relative to non-dominant airlines. While congestion pricing in the internalizing specification disproportionately benefits dominant airlines, it makes access to airports available on a more economically neutral basis. The current weight-based pricing system subsidizes non-dominant operations from dominant operations.

Columns five and six of Table 2 show the expected queue lengths for the unpriced and priced equilibria. With some exceptions, this specification has expected queues that are about the same in the unpriced case and somewhat higher in the priced case than the non-internalizing specification, so there is less reduction in queuing time as a result of pricing. Internalization of unpriced queuing delay flattens the peaks relative to non-internalizing equilibria, while the priced queues are more peaked because queuing is relatively less costly. Toronto's average queues decrease from about 3.2 to 2.6 minutes, and Vancouver's from about 4 to about 2.5 minutes. Calgary's and Montreal's average queuing decrease from about 2.5 to 1.8. These differences are less than in the non-internalizing case. While queuing levels are modest, they increase significantly with projected growth in traffic over the next decade. The current low queuing levels, particularly for the congestion pricing cases suggest that the leading Canadian airports have excess capacities for current traffic levels.

Columns seven, eight, and nine of Table 2 report the welfare and revenue changes as a result of congestion pricing in the internalizing specification. Congestion pricing recovers over $260 of delays per aircraft at Toronto, and nearly $175 at Vancouver. Calgary and Montreal save about $90 per aircraft. Social savings per aircraft and airport as shown in columns seven and nine are roughly one and a half times those in the non-internalizing specification because layover and connection time values are higher. Annual social savings at all four airports is about one hundred five million dollars, compared with seventy-two million in the non-internalizing case. Annual airport revenues from congestion pricing total about two hundred seventy-six million dollars compared with three hundred ten million in the non-internalizing specification. The unequivocal conclusion is that substantial welfare gains result from congestion pricing under either specification.

To indicate the sensitivity of these results to differences in service rates, Table 3 reports service rate elasticities of the fees, time costs, delay times, and welfare effects that appear in Tables 1 and 2. Actual airport service rates vary depending on such factors as weather conditions, mix of landings and takeoffs, mix of aircraft types, and noise regulations. Each airport's master plan [1, 8, 17, 30] reports its service rate based on the maximal number of aircraft it can land and takeoff in an hour under balanced traffic conditions. Daniel and Harback [13] calibrate the stochastic queuing model to fit estimated queuing patterns by fine tuning the reported service rates of each airport. Based on their results, the stochastic queuing model is valid for a wide variety of airport queues, but reported service rates may be slightly higher than the best parameterizations of the stochastic queuing model, causing underestimates of delays. Canadian data for estimating airport queues by time of day is not publicly available, so the service rates

cannot be fine-tuned. The (arc) elasticities in Table 3 assume ten percent reductions in service rates and should provide reasonable approximations of changes in results over the range of uncertainty about service rates.

Under congestion pricing with non internalizing behavior, the average fee, delay, time cost, and welfare gains generally decrease more than proportionately with increases in service rates. There are some anomalies among the unscheduled aircraft due to their small numbers, lower costs, greater shifting of operating times, and different operating time preferences. Under weight-based fees the direction of changes are usually the same, but the magnitude is usually smaller. Time and fee cost under weight-base pricing are especially insensitive to service rates because the large weight-based fee is fixed. For the internalizing case, the qualitative effects are similar to the above but the sensitivities of the results to the service rates are significantly smaller. It is plausible that the "true" service rate is as much as ten percent lower, meaning that the welfare gains from congestion pricing would be roughly fifteen percent higher in the non internalizing case or six percent higher in the internalizing case than reported above.

Table 4 shows the changes in fees, delays, costs, and welfare as traffic rates increase according to projections in the airports' master plans [1, 8, 17, 30]. Toronto projects thirty percent traffic growth by 2015 and nearly fifty percent by 2020 relative to the base year 2008. Calgary projects twenty percent growth by 2015 and thirty-four by 2020. Vancouver projects seventeen percent growth by 2015 and thirty-one by 2020. Montreal projects sixteen percent growth by 2015 and twenty-six by 2020. These projections are slightly above the long run trend that Transport Canada's *Aircraft Movement Statistics,* identifies. According to the Master Plans, Toronto intends to add a sixth runway around 2014, taking its capacity from 126 movements per hour to 140. Vancouver and Calgary plan additional parallel runways around 2020. The Master Plans do not include specific implementation schedules for the runway projects. Montreal apparently has no plans for additional runways within its current twenty year planning horizon, and it is not clear that the current site could accommodate an additional runway. The other airports will reach their sites' maximal capacities with the completion of one additional runway each.

If its projected traffic materializes, Toronto's current modest level of delay will more than double by 2015 under the current pricing system without a new runway. The time costs per aircraft (not including the weight-based fee) would increase by approximately two-thirds to about $1500 ($3500 internalizing). This would qualify it as moderately congested relative to US airports on the basis of queuing time, and highly congested by time costs. It is understandable that the airport would seek additional capacity under these conditions. Under congestion pricing, however, time costs per aircraft would just reach the current level in 2014 (2012 internalizing). Average queuing time would remain below four minutes, and the congestion fees would be well below current weight-based fees on average. Moreover, if the unpriced congestion level for 2014 is the correct point at which to add capacity, the airport would not reach those levels until after 2020 under congestion pricing (2020 internalizing). Queuing delays would remain around five and a half minutes (six and a half internalizing), well within levels experienced at other major

international airports. Congestion fees would average more than the current weight-based fees, but not as high as current aeronautical revenues from landing fees and terminal charges.

Projected growth rates for Vancouver, Calgary, and Montreal are lower so that they never reach the extreme congestion level projected for Toronto in 2020. Vancouver's queuing time would reach just above 3 minutes (4 minutes internalizing) and its congestion fees would reach $840 ($1200 internalizing). In no case would Calgary or Montreal reach three minutes of delay or $400 congestion fees under congestion pricing before 2020. Vancouver would only reach its current cost per aircraft in 2023 (2015 internalizing) under congestion pricing. Calgary and Montreal reach their current cost per aircraft in 2020 (Calgary 2015 and Montreal 2013 internalizing) under congestion pricing.

These results show that in addition to reducing social cost under current traffic conditions, congestion pricing produces substantial gains by using existing capacity more efficiently and postponing capital expenditures on new runways.

Policy Implications and Conclusions

As a form of marginal cost pricing, congestion pricing is a first-best solution to airport congestion. Real world approaches may not achieve all possible efficiency gains due to political feasibility constraints, but it is useful to know what effects fully internalizing congestion would have on prices, delays, traffic patterns, revenues, and welfare. The stochastic bottleneck model includes important features of airport delays that are largely absent from other airport congestion models. Modeling the trade off between layover, connection, and queuing delays is necessary to generate endogenous demand peaking caused by hub-and-spoke airlines. Accounting for changes in layover delays and connection times is also necessary to accurately quantify the effects of congestion pricing. Airlines and policymakers should consider these schedule delays along with queuing delays. The bottleneck model also has state-contingent queues, unlike many other airport congestion models in which delays are functions of contemporaneous traffic rates alone. Rapid fluctuations in traffic rates make airport queues highly dependent on the state (*i.e.*, the pdf on lengths) of queues in addition to contemporaneous traffic rates. The bottleneck model allows (nearly) continuous-time variation in traffic rates, whereas most other airport congestion models aggregate traffic by hour or half hour. Models that use half-hourly traffic data for congestion pricing or slot allocation are simply inappropriate for most airports, as the traffic patterns in Figure 1 clearly illustrate.

The stochastic bottleneck model estimates congestion prices under internalizing or non-internalizing specifications, depending on what behavioral assumptions are made about the dominant and non-dominant airlines. The reader is free to adopt either set of results. Many researchers apparently believe that it is incorrect to model dominant airlines as “naively” ignoring their self-imposed congestion. On the other hand, assuming that dominant airlines internalize their self-imposed congestion implicitly assumes that they ignore other airlines’ behavioral responses to their scheduling changes. Empirical evidence on this point is mixed, suggesting that dominant airlines behave differently at different airports. Their behavior may be influenced by the number of other dominant airlines, the strength of their hub operations, the

amount of airport capacity, the presence of slot constraints, and the proportion of dominant to non-dominant aircraft.

Recently, various proposals for slot constraint systems have gained traction with researchers and policymakers. These new proposals involve slot auctions and slot markets rather than administered slot allocations similar to those of Toronto and European airports. The results in Tables 1 and 2 indicate that Toronto's slot constraints leave unrealized forty to sixty million dollars in annual welfare gains. The IATA's *Worldwide Scheduling Guidelines* may be accurately, if uncharitably, characterized as industry administered regulation of entry and exit explicitly designed to preserve incumbent airlines' private quasi-property rights in public airports. New entrants may only obtain slots from new capacity or from slots forfeited due to disuse. Slots are not purchased or sold. Initial slot allocations are granted on the basis of historical use at the beginning of semiannual scheduling conferences. Nine hundred airline and airport representatives meet for three days to conduct bilateral negotiations for slot exchanges. It is difficult to know whether replacing this system with slot auctions and/or slot markets would reduce the transaction costs of allocating slots. The current system illustrates that airlines need to optimize complex combinations of slots to support their flight schedules. Combinatorial auction mechanisms for contingent bidding on bundles of slots may be even more costly than the IATA's approach. Aside from the problems of conducting the auction, the proposals also fail to address the problem of demand peaking within the slot windows.

Proponents of slot auctions and markets argue that optimal congestion prices are difficult to set, whereas auctions and markets reveal the participants' valuations of slots. The difficulty of setting congestion prices is overstated. Given the expected arrival rates by time of day—obtained by averaging arrival rates by time period (minute) over a number of days—the Appendix provides analytical expressions for congestion fees that depend on parameters for: the airport's service rate; the value of layover, connection, and queuing time; and the preferred operating times. Airlines will adjust their aircraft operating times in response to fees. The economist needs to anticipate the equilibrium traffic patterns to predict what the congestion fee schedules and social gains will look like, but it is not necessary for the airport authority to compute the equilibrium to determine the external costs. The airport authority can simply use the observed traffic patterns that emerge in response to pricing as they slowly phase in congestion pricing. Congestion pricing avoids the administrative overhead of allocating slots. Airlines are free to fly anywhere at any time, provided they are willing to pay the fee and experience the delay. Markets will always clear as queues adjust to equilibrate supply and demand, just as they do currently at nearly every US airport. As congestion fees approach the full external costs as calculated in the stochastic queuing model, the traffic rate should approach the equilibrium pattern predicted by the bottleneck model, just as the current traffic rates follow the no-fee bottleneck equilibrium. If the model is imperfect, non-optimal queues will still equilibrate supply and demand, and the traffic rates will converge to a suboptimal pattern that probably will nevertheless enhance welfare. Pricing should be reevaluated shortly after implementation to look for inefficiencies. Many policy makers apparently dislike fee schedules that fluctuate rapidly and continuously.

This model and the empirical evidence on airport traffic patterns, however, suggests that the only real policy choice is between rapidly fluctuating congestion fees (that generate substantial revenues) or rapidly fluctuating queuing delays (with significant welfare losses).

Based on the specification tests in Daniel and Harback [12] and the welfare results presented in Tables 1, the author argues that there is a strong case for imposing non-internalizing fee structures as shown in the bottom panels of Figures 3 to 6 on aircraft operations at Toronto, Vancouver, Calgary and Montreal. For readers who believe the internalizing specification, Table 2 makes a strong case for imposing internalizing fee structures as shown in the bottom panels of Figures 7 to 10.

REFERENCES

[1] Aeroports De Montreal*, Strategic Plan Summary 2008::2012, 2008.*

[2] Air Transport Research Society, *The ATRS Global Airport Benchmarking Report 2005.*

[3] Arnott, R., A. de Palma, and R. Lindsey. Economics of a Bottleneck, *Journal of Urban Economics* 24(1) (1990) 111-126.

[4] _____, _____, and _____, A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand, *American Economic Review* 83(1) (1993) 161-179.

[5] Borins, Sandford F., Pricing and investment in a transportation network: the case of Toronto Airport, *Canadian Journal of Economics*, XI no. 4, Nov. 1978.

[6] Brueckner, Jan K., Airport Congestion When Carriers Have Market Power, *American Economic Review* 92(5) (2002) 1357-1375.

[7] _____ and Kurt Van Dender, Atomistic congestion tolls at concentrated airports? Seeking a unified view in the internalization debate, *Journal of Urban Economics*, Forthcoming, (2008).

[8] Calgary International Airport, *Master Plan*, 2004.

[9] Carlin, A. and R. Park (1970): "Marginal Cost Pricing of Airport Runway Capacity," *The American Economic Review*, 60 (June), 310-319.

[10] Daniel, Joseph I., Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues, *Econometrica* 63(2) (1995) 327-370.

[11] _____, "Distributional Consequences of Airport Congestion Pricing," *Journal of Urban Economics* 50:2, 2001, 230-258.

[12] _____ and Katherine Harback, "(When) do hub airlines internalize their self-imposed delays?"*Journal of Urban Economics* 63 (2008) 583-612.

[13] _____ and Katherine Harback, "Pricing the Major US Airports," University of Delaware Working Paper (2008).

[14] _____ and Munish Pahwa, Comparison of Three Empirical Models of Airport Congestion Pricing, *Journal of Urban Economics* 47(1) (2000) 1-38.

[15] FAA, *Economic Values for FAA Investment and Regulatory Decisions, a Guide,* 2005.

[16] Gillen, D., Henriksson, L., et al., "Airport Financing, Costing, Pricing and Performance," Final Report to the Canadian Transportation Act Review Committee, April 2001.

[17] Greater Toronto Airport Authority, *Taking Flight – The Airport Master Plan 2008–2030* , 2007.

[18] Henderson, J. V. (1985): *Economic Theory and the Cities*, 2nd ed. New York, N.Y. Academic Press.

[19] International Air Transport Association, *Worldwide Scheduling Guidelines, 12th Edition, 2005.*

[20] Jones, I., Holder, S., et al., "Study to Assess the Effects of Different Slot Allocation Schemes, A Report for the European Commission," NERA Economic Consulting, January 2004.

[21] Koopman, B. (1972): "Air-Terminal Queues under Time Dependent Conditions," *Operations Research*, 20, 1089 - 1114.

[22] Mayer, C and Todd Sinai, Network Effects, Congestion Externalities, and Air Traffic Delays: Or Why Not All Delays Are Evil, *American Economic Review* 93(4) (2003) 1194-1215.

[23] Morrison, S. (1983): "Estimation of Long-Run Prices and Investment Levels for Airport Runways," *Journal of Transportation Economics*, 1, 103 - 130.

[24] Morrison, S., and C. Winston (1989): "Enhancing the Performance of the Deregulated Air Transportation System," *Brookings Papers on Economic Activity, Microeconomics*, 61-123.

[25] Omosigho, S. E. and D. J. Worthington (1988): "An approximation of known accuracy for single server queues with inhomogeneous arrival rate and continuous service time distribution," *European Journal of Operational Research* 33:3, 304-313

[26] Oum, Tae Hoon and Zhang, Yimin. "Airport Pricing: Congestion Tolls, Lumpy Investment, and Cost Recovery," *Journal of Public Economics*, Dec. 1990, Vol. 43(3), 353-375.

[27] Vickrey, William S., Congestion Theory and Transport Investment, *American Economic Review* (Papers and Proceedings) 59(2) (1969) 251-60.

[28] Treasury Board of Canada Secretariat, *Benefit Cost Analysis Guide,* Ottawa, 1998.

[29] Transport Canada, Aircraft Movement Statistics, Annual Report 2005.

[30] Vancouver International Airport Authority *Draft 20-Year Master Plan*, 2006.

[31] Verhoef, Erik T. "Congestion Pricing, Slot Sales and Slot Trading in Aviation," Tinbergen Institute Discussion Paper TI 2008-030/3.

[32] Whalen, W. T., Carlton, D.W., et al., "Proposal For A Market-Based Solution to Airport Delays," U.S. Department of Justice, Economic Analysis Group Discussion Paper 07-14, October 2007.

Appendix A

M(t)/d/s/K queuing systems have Poisson arrivals with time varying rates (λ_t arrivals per minute), deterministic service intervals (*d* operations per minute), multiple servers (*s* parallel runways), and maximum queue sizes (no more than *K* aircraft awaiting service). A state vector ($\boldsymbol{q}_t$) has elements, $[q_{0t}, q_{1t}, q_{2t}, \ldots, q_{Kt}]$, representing the probabilities of the queue having lengths *0, 1, 2,… K* at time *t*. Let *w(k)* be the waiting time required for aircraft joining the queue when its length is *k.* Expected queuing cost at time *t* is the product of the probabilities, q_{kt}, and the waiting time in the queue, *w(k),* summed over all *k*.. The expected direct queuing cost for an aircraft actually arriving during period *t* is $Q_t = \Sigma_{k=0}^{K} q_{k,t}\, w(k)$.

The most-preferred service-completion time of a dominant aircraft *a* is τ_a, the start or end of the passenger exchange period. The aircraft arriving for service at *t* when the queue is of length *k* actually completes service at time *t+w(k)*. Layover time is the waiting time between the early service-completion time and the most-preferred service-completion time. If aircraft *a* completes service before τ_a, (i.e., $t+w(k) \le \tau_a$) then its layover time is $\tau_a - t - w(k)$. Reduced connection time is the time between the most-preferred service-completion time and the late service-completion time. If an aircraft completes service after τ_a, then its connection time is reduced by $t+w(k)-\tau_a$. To calculate expected layover and reduced-connection time at time *t*, simply weight layover and reduced-connection times by the probability q_{kt} that the queue is of length *k* at time *t* and sum over the range of *k* for which aircraft are early or late. The direct layover time of aircraft actually joining queue at time *t* is $C_t = \Sigma_{k=0}^{K} q_{k,t}\, max[0, \tau_a - t - w(k)]$. The direct reduced connection time of aircraft actually joining the queue at time *t* is $L_t = \Sigma_{k=0}^{K} q_{k,t}\, max[0, t+w(k)-\tau_a]$.

The queuing system parameters (*s*, *d*, and *K*) and the time-varying traffic rates (λ_t) fully determine the transition matrices ($\boldsymbol{T}_t$) for each service period:

A.1 $\boldsymbol{T}_t =$

$$
\begin{vmatrix}
\frac{(\lambda_t)^0(e)^{-\lambda_t}}{0!} & \cdots & \frac{(\lambda_t)^0(e)^{-\lambda_t}}{0!} & 0 & 0 & 0 \cdots & 0 \\
\frac{(\lambda_t)^1(e)^{-\lambda_t}}{1!} & \cdots & \frac{(\lambda_t)^1(e)^{-\lambda_t}}{1!} & \frac{(\lambda_t)^0(e)^{-\lambda_t}}{0!} & 0 & 0 \cdots & 0 \\
\frac{(\lambda_t)^2(e)^{-\lambda_t}}{2!} & \cdots & \frac{(\lambda_t)^2(e)^{-\lambda_t}}{2!} & \frac{(\lambda_t)^1(e)^{-\lambda_t}}{1!} & \frac{(\lambda_t)^0(e)^{-\lambda_t}}{0!} & 0 \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \; \vdots & \vdots \\
\frac{(\lambda_t)^{K-1}(e)^{-\lambda_t}}{(K-1)!} & \cdots & \frac{(\lambda_t)^{K-1}(e)^{-\lambda_t}}{(K-1)!} & \frac{(\lambda_t)^{K-2}(e)^{-\lambda_t}}{(K-2)!} & \frac{(\lambda_t)^{K-3}(e)^{-\lambda_t}}{(K-3)!} \cdots & \cdots & \frac{(\lambda_t)^{s-1}(e)^{-\lambda_t}}{(s-1)!} \\
1 - \sum_{r=0}^{K-1} \frac{(\lambda t)^r(e)^{-\lambda_t}}{r!} & \cdots & 1 - \sum_{r=0}^{K-1} \frac{(\lambda t)^r(e)^{-\lambda_t}}{r!} & 1 - \sum_{r=0}^{K-2} \frac{(\lambda t)^r(e)^{-\lambda_t}}{r!} & 1 - \sum_{r=0}^{K-3} \frac{(\lambda t)^r(e)^{-\lambda_t}}{r!} \cdots & \cdots & 1 - \frac{(\lambda_t)^{s-1}e)^{-\lambda_t}}{(s-1)}
\end{vmatrix}
$$

<---------- *s+1 terms* ----------><----------*K-s terms*----------->

Multiplying the state vector by the period's transition matrix gives the next period's state vector ($\boldsymbol{q}_{t+1} = \boldsymbol{T}_t \boldsymbol{q}_t$). Iteratively multiplying the state vector by the transition matrices determines state of the queue in every subsequent period; i.e., $\boldsymbol{q}_{t+i} = \boldsymbol{T}_{t+i}\, \boldsymbol{T}_{t+i-1} \ldots\, \boldsymbol{T}_t\, \boldsymbol{q}_t$.

Multiplying each period's expected queue lengths by the probability that a given aircraft actually arrives during that period (from the distribution of actual about intended operating times $\boldsymbol{s}_{t,i}$ and summing over all the periods in which it potentially operates determines the expected direct queuing delay of an arrival scheduled at period *t:*

A.2 $E[Q_t] = \Sigma_i s_{t,i} \Sigma_{k=0}^{K} q_{k,(t+i)} w(k).$

Similarly, the expected direct layover and reduced connection times of an arrival scheduled at period *t* are:

A.3 $E[L_t] = \Sigma_i s_{t,i} \Sigma q_{k,(t+i)} \max[0, \tau_a-(t+i)-w(k)].$

A.4 $E[C_t] = \Sigma_i s_{t,i} \Sigma q_{k,t+i} \max[0, (t+i)+w(k)-\tau_a]$

The marginal effect of a current arrival on queuing in the next period is determined by pre-multiplying the state vector $\boldsymbol{q}_t$ by a transition matrix $\boldsymbol{D}_t$ obtained by differentiating the current queuing transition matrix, $\boldsymbol{T}_t$, element by element with respect to the current arrival rate; i.e., $\boldsymbol{\delta}_{t,t+1} = \boldsymbol{D}_t \boldsymbol{q}_t$ where:

A.5 $\boldsymbol{D}_t =$

$$\begin{vmatrix}
\frac{0(\lambda_t)^{-1}-(\lambda_t)^0}{0!} & \cdots & \frac{0(\lambda_t)^{-1}-(\lambda_t)^0}{0!} & 0 & 0 & 0 & \cdots & 0 \\
\frac{1(\lambda_t)^0-(\lambda_t)^1}{1!} & \cdots & \frac{1(\lambda_t)^0-(\lambda_t)^1}{1!} & \frac{0(\lambda_t)^{-1}-(\lambda_t)^0}{0!} & 0 & 0 & \cdots & 0 \\
\frac{2(\lambda_t)^1-(\lambda_t)^2}{2!} & \cdots & \frac{2(\lambda_t)^1-(\lambda_t)^2}{2!} & \frac{1(\lambda_t)^0-(\lambda_t)^1}{1!} & \frac{0(\lambda_t)^{-1}-(\lambda_t)^0}{0!} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{(K-1)(\lambda_t)^{K-2}-(\lambda_t)^{K-1}}{2!} & \cdots & \frac{(K-1)(\lambda_t)^{K-2}-(\lambda_t)^{K-1}}{2!} & \frac{(K-2)(\lambda_t)^{K-3}-(\lambda_t)^{K-2}}{(K-2)!} & \frac{(K-3)(\lambda_t)^{K-4}-(\lambda_t)^{K-3}}{(K-3)!} & \cdots & \cdots & \frac{(s-1)(\lambda_t)^{s-2}-(\lambda_t)^{s-1}}{(K-3)!} \\
\sum_{r=0}^{K-1} \frac{(\lambda_t)^r-r(\lambda_t)^{r-1}}{r!} & \cdots & \sum_{r=0}^{K-1} \frac{(\lambda_t)^r-r(\lambda_t)^{r-1}}{r!} & \sum_{r=0}^{K-2} \frac{(\lambda_t)^r-r(\lambda_t)^{r-1}}{r!} & \sum_{r=0}^{K-3} \frac{(\lambda_t)^r-r(\lambda_t)^{r-1}}{r!} & \cdots & \cdots & \sum_{r=0}^{s-3} \frac{(\lambda_t)^r-r(\lambda_t)^{r-1}}{r!}
\end{vmatrix}$$

<---------- *s+1 terms* ----------><----------*K-s terms*----------->

The vector $\boldsymbol{\delta}_{t,t+1}$ gives the rate of change in probabilities that the queue is of lengths *0* to *k* in the next period with respect to the current arrival rate, λ_t. Iteratively multiplying this vector by the transition matrices determines the vectors that specify the marginal changes in the probability distributions on queue lengths for all subsequent periods ($\boldsymbol{\delta}_{t,t+i} = \boldsymbol{T}_{t+i}\, \boldsymbol{T}_{t+i-1} \ldots \boldsymbol{T}_{t+1} \boldsymbol{D}_t\, \boldsymbol{q}_t$).

Let the set $\boldsymbol{A}$ contain the intended operating times t_α of all aircraft, and the set $\boldsymbol{H}$ contain only the intended operating times t_η of non dominant airlines' aircraft. The expected marginal external queuing delay that an aircraft intended to operate at time t_a imposes on an aircraft intended to arrive at period t_x is $\Sigma_i s_{t_x,i} \Sigma_{k=0}^{K} \delta_{k,t_a,t_x+i} w(k)$. Summing over all aircraft operating times in sets $\boldsymbol{A}$ or $\boldsymbol{H}$ determines the expected marginal queuing delay that an actual arrival at period *t* imposes externally, $Q'_{t_a} = \Sigma_{x \epsilon X} \Sigma_i s_{t_x,i} \Sigma_{k=0}^{K} \delta_{k,t_a,t_x+i} w(k)$, where the set $\boldsymbol{X}$ represents either $\boldsymbol{A}$ *or* $\boldsymbol{H}$. Similarly, the expected marginal layover and reduced connection times that an actual arrival at period t_a imposes externally are $L'_{t_a} = \Sigma_{x \epsilon X} \Sigma_i s_{t_x,i} \Sigma_{k=0}^{K} \delta_{k,t_a,t_x+i} \max[0, \tau_a-(t_x+i)-w(k)]$, and $C'_{t_a} = \Sigma_{x \epsilon X} \Sigma_i s_{t_x,i} \Sigma_{k=0}^{K} \delta_{k,t_a,t_x+i} \max[0, (t+i)+w(k)-\tau_a]$.

The expected marginal external queuing delays for aircraft intended to operate at time t is the sum of the probabilities of actually arriving in periods $s_{t,,i}$ times these expected marginal delays at time $t_{.}+i$:

A.6 $E[Q'_{t_a}]=\sum_i s_{t_a,t_a+i} \sum_{x\in X} \sum_i s_{t_x,i} \sum_{k=0}^{K} \delta_{k,t,t_x+i}\, w(k)$

A.7 $E[L'_{t_a}]=\sum_i s_{t_a,t_a+i} \sum_{x\in X} \sum_i s_{t_x,i} \sum_{k=0}^{K} \delta_{k,t,t_x+i}\, max[0,\ \tau_a-(t_x+i)-w(k)]$, and

A.8 $E[C'_{t_a}]=\sum_i s_{t_a,t_a+i} \sum_{x\in X} \sum_i s_{t_x,i} \sum_{k=0}^{K} \delta_{k,t,t_x+i}\, max[0,\ (t+i)+w(k)-\tau_a]$.

Note that all these calculations give exact analytical expressions for marginal delay times, and their derivations are independently verifiable under the assumptions of the model. The values of these expressions, however, depend on the endogenous values of the intended operating times t_a of all aircraft.

The bottleneck equilibria are Nash equilibria in which the players are individual airlines denoted by elements n in the set of all airlines $\boldsymbol{N}$; the strategies are scalars or vectors of intended operating times $\boldsymbol{t}_n=[t_1,\ t_2,\ \ldots,t_{m_n}]$ where $m_n\geq 1$ is the number of aircraft operated by airline n; and each $\boldsymbol{t}_n$ is a best response that minimizes the airline's total operating costs subject to the schedule choices $\boldsymbol{t}_o$ of all the other airlines $o\in\boldsymbol{N}$. The best responses depend on the behavioral assumptions about internalization and whatever fee schedule the airport imposes. Let v_Q, v_C, and v_L denote the value of queuing, connection, and layover times.

A fully atomistic, no-fee bottleneck equilibrium has airlines choosing schedule times that minimize direct aircraft costs:

A.9 $t_j=arg\ min\ \{v_Q\, E[Q_{t_j}]+v_C\, E[C_{tj}]+v_L\, E[L_{tj}]\},\ \forall t_j\in\boldsymbol{t}_n$ *and* $n\in\boldsymbol{N}$ *and given all* $t_k\in\boldsymbol{t}_o$, *where* $o\in N$.

A fully atomistic, priced bottleneck equilibrium has airlines choosing schedule times that minimize the sum of direct aircraft costs plus the fee $F[t_j\ ;\ \boldsymbol{t}_{\sim j}]=v_Q\, E[Q'_{t_j}]+v_C\, E[C'_{tj}]+v_L\, E[L'_{tj}]$:

A.10 $t_j=arg\ min\ \{v_Q\, E[Q_{t_j}]+v_C\, E[C_{tj}]+v_L\, E[L_{tj}]+F[t_j\ ;\ \boldsymbol{t}_{\sim j}]\}$,

$\forall t_j\in\boldsymbol{t}_n$ *and* $n\in\boldsymbol{N}$ *and given all* $t_k\in\boldsymbol{t}_o$, *where* $o\in N$.

A dominant-non dominant, no-fee bottleneck equilibrium has dominant airlines choosing schedule times that minimize the sum of direct and indirect costs of its own aircraft:

A.11 $t_j=arg\ min\ \{v_Q\, E[Q_{t_j}]+v_C\, E[C_{tj}]+v_L\, E[L_{tj}]+v_Q\, E[Q'_{t_j}]+v_C\, E[C'_{tj}]+v_L\, E[L'_{tj}]$,

$\forall t_j\in\boldsymbol{t}_n$ *and* $n\in\boldsymbol{N}$ *and given all* $t_k\in\boldsymbol{t}_o$, *where* $o\in N$;

and non dominant airlines choose their t_j as specified in A.9.

A dominant-non dominant, ideally priced bottleneck equilibrium has dominant airlines choosing schedule times that minimize the sum of direct and indirect costs of its own aircraft plus the fee that includes the indirect costs imposed on other airlines' aircraft, $F[t_j;\boldsymbol{t}_{\sim j}]=v_Q E[Q'_{t_j}]+v_C E[C'_{tj}]+v_L E[L'_{tj}]$,:

A.12 $t_j=arg\ min\ \{v_Q\, E[Q_{t_j}]+v_C\, E[C_{tj}]+v_L\, E[L_{tj}]+v_Q\, E[Q'_{t_j}]+v_C\, E[C'_{tj}]+v_L\, E[L'_{tj}]+F[t_j;\boldsymbol{t}_{\sim j}]$,

$\forall t_j\in\boldsymbol{t}_n$ *and* $n\in\boldsymbol{N}$ *and given all* $t_k\in\boldsymbol{t}_o$, *where* $o\in N$;

and non dominant airlines choose their t_j as specified in A.9.

[1] The author gratefully acknowledges the support of this research by a grant from the University of Delaware's Lerner College of Business and Economics.

[2] Toronto follows the International Air Transit Association's (IATA) *Worldwide Scheduling Guidelines* (*WSG*) [19] for allocating slots. The basic principles of the *WSG* provide for allocation based on historical usage subject to use-it-or-lose-it rules, allocation of a fraction of new or unused slots to new entrants, and bilateral one-for-one trading of existing slots. The guidelines call for flexibility in permitting deviations of actual operating times from slot times and do not specify penalties for excessive deviations other than possibly losing the slot.

[3] Daniel and Harback [12] found similar peaking at slot controlled airports in the U.S.

[4] These values for layover and connection time for arrivals and departures are 0.0, 0.0, 0.0, and 0.0 repectively.

[5] Similar graphs for the departures are shown in Tables A.1 and A.2 of the appendix. The equilibria include unscheduled traffic, but these graphs only include scheduled traffic for direct comparison with scheduled data. The scheduled traffic is distributed around the schedule time using the distribution of actual arrivals around mean operating times from MSP airport. The traffic patterns are not very sensitive to the observed variations in these distributions among US airports.

[6] In 2005-2006 it was the highest in the world. Air Transport Research Society, *The ATRS Global Airport Benchmarking Report 2005* [2].

[7] Management's discussion and analysis and Consolidated Financial Statements of the GREATER TORONTO AIRPORTS AUTHORITY December 31, 2007, pp. 7, 10.

[8] VANCOUVER AIRPORT AUTHORITY, *TARIFF OF FEES AND CHARGES*, effective January 1, 2008.

[9] The Calgary Airport Authority, *TARIFF OF AVIATION FEES*, as at January 1, 2007.

[10] Landing fee revenues from *Aeroports de Montreal Annual Report 2007* [1], Management's discussion and analysis for the year ended December 31, 2007, p. 27. MTOW estimated from Transport Canada's, *Aircraft Movement Statistics2005,* p. 27 [29]

[11] Airport Improvement Fees, http://www.aircanada.com/shared/en/common/flights/pop_surcharge.html

Table 1--Time, Costs, and Social Savings

	Number of Arrivals	Size of Arrivals (1000 kg)	Average Weight-Based Fee	Average Congestion Fee	Time & Fee Costs per Aircraft, Weight-Based Fee	Time & Fee Costs per AIrcraft, Congestion Pricing	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Social Savings per Aircraft	Change in Daily Cost per Operator	Change in Daily Revenues	Net Gain per Airport
Toronto												$117,984
Air Canada & Jazz	364	104	$3,484	$892	$4,335	$1,570	3.96	2.36	$172	-$1,006,407	-$943,661	
Other Scheduled	262	125	$4,206	$868	$5,349	$1,800	2.88	2.30	$211	-$929,959	-$874,693	
Unscheduled	30	12	$406	$504	$499	$598	3.06	1.81	-$1	$2,975	$2,947	
Vancouver												$50,226
Air Canada	166	100	$486	$460	$1,117	$933	3.83	2.09	$157	-$30,444	-$4,365	
Other Other Scheduled	164	100	$486	$455	$1,106	$922	3.80	2.09	$153	-$30,154	-$5,111	
Unscheduled commercial	92	25	$103	$184	$229	$313	2.21	1.76	-$2	$7,674	$7,499	
Unscheduled private	24	10	$32	$152	$87	$237	2.30	1.75	-$30	$3,593	$2,872	
Calgary												$15,066
Air Canada & Jazz	113	80	$463	$222	$802	$477	2.90	1.60	$84	-$36,678	-$27,176	
Westjet	74	68	$392	$193	$663	$411	2.70	1.59	$52	-$18,602	-$14,746	
Other Scheduled	45	125	$923	$184	$1,415	$602	2.05	1.47	$75	-$36,601	-$33,226	
Unscheduled commercial	45	25	$119	$95	$218	$213	1.87	1.35	-$20	-$202	-$1,088	
Unscheduled private	36	10	$41	$63	$81	$124	1.86	1.29	-$21	$1,570	$790	
Montreal												$14,495
Air Canada & Jazz	110	83	$1,073	$265	$1,418	$526	3.07	1.72	$84	-$98,157	-$88,867	
Other Scheduled	99	125	$1,625	$210	$2,104	$636	1.91	1.53	$53	-$145,321	-$140,116	
Unscheduled commercial	47	25	$325	$125	$421	$219	1.82	1.39	$3	-$9,421	-$9,299	
Unscheduled private	30	10	$130	$104	$169	$175	1.64	1.36	-$32	$177	-$786	

Table 2--Time, Costs, and Social Savings

	Average Weight-Based Fee	Average Congestion Fee	Time & Fee Costs per Aircraft, Weight-Based Fee	Time & Fee Costs per AIrcraft, Congestion Pricing	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Social Savings per Aircraft	Change in Daily Cost per Operator	Change in Daily Revenues	Net Gain per Airport
Toronto										$170,218
Air Canada & Jazz	$3,484	$481	$5,524	$2,273	3.25	2.76	$248	-$1,183,341	-$1,093,156	
Other Scheduled	$4,206	$1,022	$6,313	$2,806	3.09	2.54	$323	-$918,776	-$834,230	
Unscheduled	$406	$457	$548	$750	3.08	1.83	-$150	$6,060	$1,547	
Vancouver										$72,769
Air Canada	$486	$385	$2,129	$1,643	3.58	2.80	$385	-$80,648	-$16,711	
Other Other Scheduled	$486	$734	$1,443	$1,591	5.17	2.76	$100	$24,185	$40,647	
Unscheduled commercial	$103	$185	$353	$494	2.99	1.86	-$58	$12,352	$7,268	
Unscheduled private	$32	$139	$132	$337	3.22	1.75	-$98	$5,327	$2,782	
Calgary										$20,971
Air Canada & Jazz	$463	$131	$1,148	$696	2.73	2.00	$121	-$51,044	-$37,428	
Westjet	$392	$145	$907	$586	2.92	2.00	$73	-$23,721	-$18,322	
Other Scheduled	$923	$232	$1,703	$917	2.39	1.73	$96	-$35,372	-$31,065	
Unscheduled commercial	$119	$115	$279	$294	2.11	1.47	-$19	$644	-$202	
Unscheduled private	$41	$61	$105	$167	2.09	1.23	-$42	$2,226	$719	
Montreal										$22,690
Air Canada & Jazz	$1,073	$166	$1,847	$805	2.93	2.21	$134	-$114,519	-$99,832	
Other Scheduled	$1,625	$252	$2,478	$1,018	2.23	1.75	$87	-$144,516	-$135,893	
Unscheduled commercial	$325	$146	$498	$312	2.10	1.60	$7	-$8,550	-$8,215	
Unscheduled private	$130	$94	$199	$195	2.07	1.32	-$32	-$113	-$1,067	

Table 3--Service Rate Elasticities, Non Internalizing Case

	Average Congestion Fee	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Time & Fee Costs per Aircraft, Weight-Based Fee	Time & Fee Costs per AIrcraft, Congestion Pricing	Social Savings per Aircraft	Net Gain per Airport
Toronto							-1.79
Air Canada & Jazz	-2.88	-1.16	-1.62	-0.17	-2.14	-0.41	
Other Scheduled	-3.21	-1.93	-1.84	-0.07	-1.69	-2.77	
Unscheduled	-3.73	-1.67	-2.07	-0.16	-2.89	-10.15	
Vancouver							-1.24
Air Canada	-2.28	-0.92	-1.35	-0.33	-1.45	-0.90	
Other Other Scheduled	-2.23	-0.94	-1.35	-0.34	-1.43	-0.95	
Unscheduled commercial	-3.50	-1.65	-2.17	-0.24	-2.21	-11.36	
Unscheduled private	-4.26	-1.74	-2.46	-0.33	-2.82	28.89	
Calgary							-1.37
Air Canada & Jazz	-1.41	-0.75	-0.91	-0.21	-0.87	-0.87	
Westjet	-1.32	-0.84	-0.88	-0.21	-0.83	-1.10	
Other Scheduled	-1.69	-1.19	-1.09	-0.05	-0.53	-1.26	
Unscheduled commercial	-2.36	-1.31	-1.77	-0.02	-1.10	1.97	
Unscheduled private	-2.89	-1.35	-1.65	-0.07	-1.13	9.54	
Montreal							-1.34
Air Canada & Jazz	-1.91	-0.83	-1.11	-0.12	-1.21	-0.83	
Other Scheduled	-2.19	-1.54	-1.38	0.01	-0.69	-1.62	
Unscheduled commercial	-2.34	-1.46	-1.75	0.03	-1.38	-4.42	
Unscheduled private	-2.56	-1.71	-1.86	-0.02	0.14	-64.22	

Table 4--Service Rate Elasticities, Internalizing Case

	Average Congestion Fee	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Time & Fee Costs per Aircraft, Weight-Based Fee	Time & Fee Costs per AIrcraft, Congestion Pricing	Social Savings per Aircraft	Net Gain per Airport
Toronto							-0.68
Air Canada & Jazz	-1.21	-0.61	-0.65	-0.18	-0.61	-0.78	
Other Scheduled	-1.32	-0.89	-0.94	0.01	-0.46	-0.41	
Unscheduled	-1.66	-0.96	-1.17	0.01	-0.68	3.10	
Vancouver							-0.51
Air Canada	-0.56	-0.30	-0.37	-0.17	-0.28	-0.30	
Other Other Scheduled	-0.49	-0.31	-0.35	-0.08	-0.27	-0.54	
Unscheduled commercial	-1.53	-0.73	-1.31	0.08	-0.45	1.43	
Unscheduled private	-1.39	-0.69	-1.32	0.02	-0.25	1.61	
Calgary							-0.39
Air Canada & Jazz	-0.49	-0.26	-0.26	-0.11	-0.18	-0.50	
Westjet	-0.61	-0.52	-0.39	-0.14	-0.27	-0.75	
Other Scheduled	-0.66	-0.65	-0.47	0.04	-0.13	0.32	
Unscheduled commercial	-0.77	-0.68	-0.78	0.09	-0.41	-2.23	
Unscheduled private	-1.39	-0.74	-1.29	0.07	-0.41	0.51	
Montreal							-0.80
Air Canada & Jazz	-0.75	-0.56	-0.34	-0.13	-0.29	-0.92	
Other Scheduled	-0.93	-0.92	-0.65	0.07	-0.03	-0.64	
Unscheduled commercial	-0.65	-0.87	-0.52	0.08	-0.36	26.80	
Unscheduled private	-1.17	-1.03	-0.99	0.07	-0.31	1.83	

Table 5--Traffic Growth, Delay, Costs, and Social Savings (Non Internalizing)

	Year	Number of Aircraft	Average Congestion Fee	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Social Costs per Aircraft, Weight-Based Fee	Social Costs per AIrcraft, Congestion Pricing	Social Savings per Aircraft	Daily Congestion Fee Revenues	Social Savings
Toronto	2008	656	864	3.5	2.3	933	753	180	567,103	117,984
1.30	2015	823	2,579	7.1	4.0	1,511	1,064	447	2,199,373	381,354
1.49	2020	931	4,663	12.7	5.3	2,428	1,341	1,086	4,557,699	1,061,816
Vancouver	2008	446	385	3.4	2.0	492	379	113	171,506	50,226
1.17	2015	500	532	4.3	2.4	609	457	152	277,796	79,314
1.31	2020	548	840	5.8	3.1	779	526	253	490,836	147,794
Calgary	2008	314	173	2.5	1.5	276	227	48	54,274	15,066
1.20	2015	356	253	3.4	1.9	348	271	77	95,187	28,958
1.34	2020	388	330	3.9	2.3	399	302	97	138,839	40,618
Montreal	2008	285	206	2.3	1.6	319	271	48	58,859	13,654
1.16	2015	316	279	2.9	1.8	370	310	60	92,242	19,764
1.26	2020	342	390	3.7	2.3	441	345	96	140,246	34,449

Table 6--Traffic Growth, Delay, Costs, and Social Savings (Internalizing)

	Year	Number of Aircraft	Average Congestion Fee	Average Delay, Weight-Based Fee	Average Delay, Congestion Pricing	Social Costs per Aircraft, Weight-Based Fee	Social Costs per AIrcraft, Congestion Pricing	Social Savings per Aircraft	Daily Congestion Fee Revenues	Social Savings
Toronto	2008	656	696	3.2	2.6	1,980	1,720	259	456,671	170,218
1.30	2015	836	2,476	5.9	4.3	3,472	2,217	1,255	2,111,648	1,070,564
1.49	2020	944	5,437	9.3	6.4	5,769	3,170	2,599	5,314,538	2,540,466
Vancouver	2008	444	460	4.0	2.5	1,024	860	164	204,246	72,769
1.17	2015	512	677	5.0	3.2	1,253	1,018	235	351,563	122,045
1.31	2020	586	1,188	6.6	4.3	1,596	1,215	380	691,167	221,208
Calgary	2008	313	139	2.6	1.8	512	445	67	43,382	20,971
1.20	2015	369	208	3.3	2.2	624	523	101	78,287	37,952
1.34	2020	422	271	3.7	2.6	692	564	128	113,493	53,488
Montreal	2008	285	185	2.5	1.9	630	550	80	52,758	22,690
1.16	2015	320	258	3.0	2.2	722	618	104	85,314	34,400
1.26	2020	370	375	3.7	2.7	804	661	143	134,599	51,274

Figure 1--Comparison of Scheduled Arrivals and Modelled Non Internalizing Arrivals

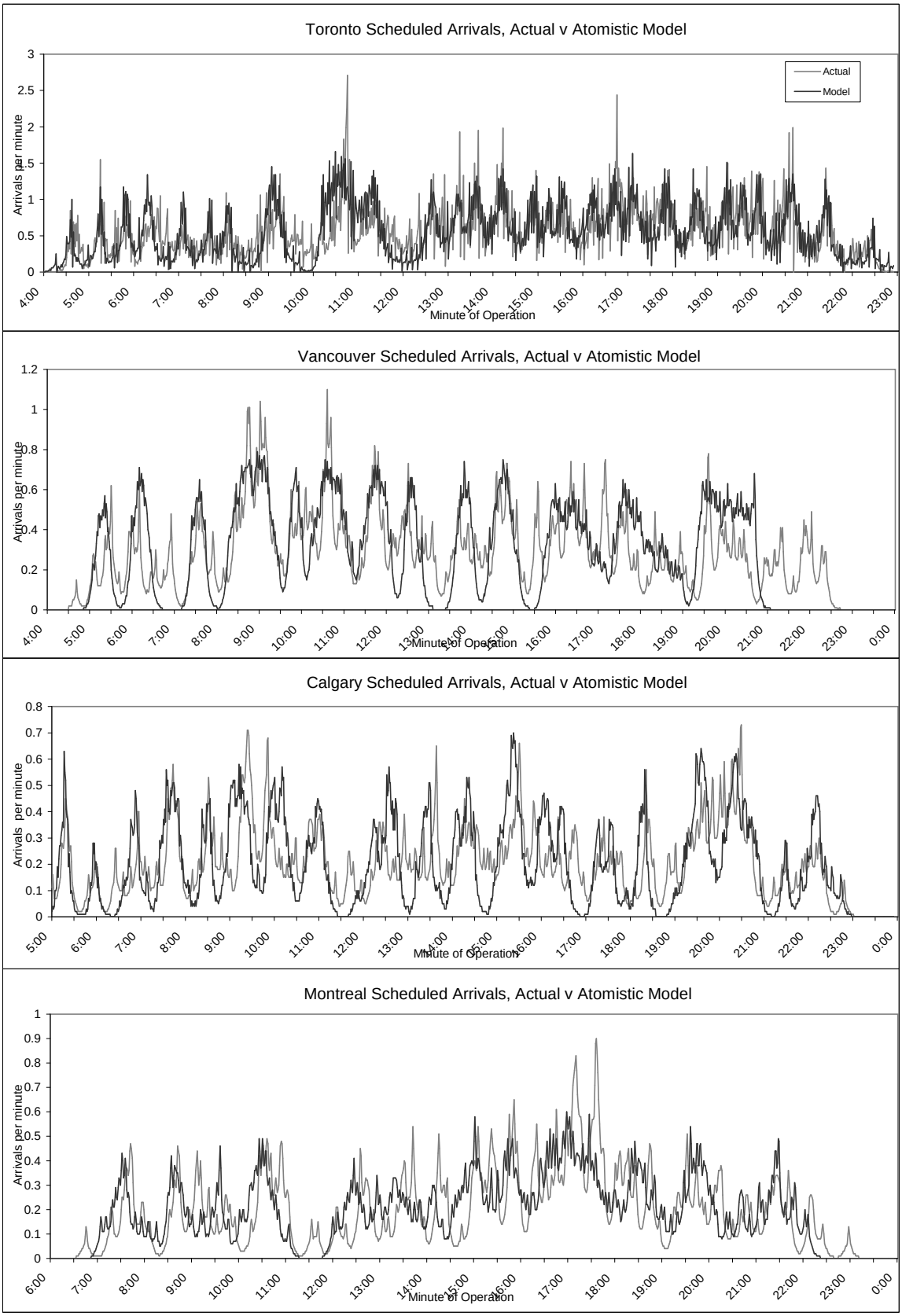

Figure 2--Comparison of Scheduled Arrivals and Modelled Internalizing Arrivals

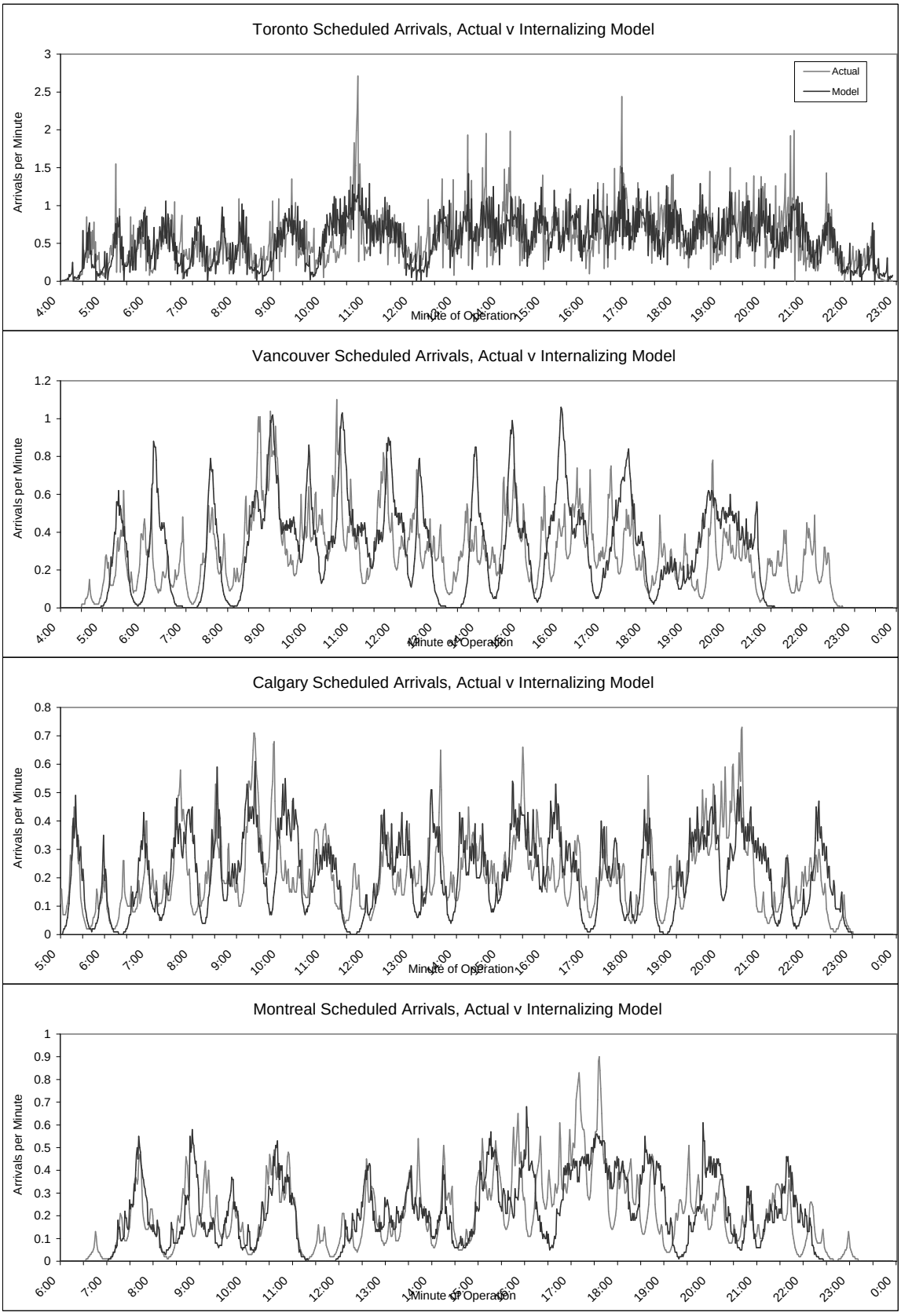

Figure 3--Non Internalizing Toronto Model, Unpriced and Priced Arrivals, Delays, and Congestion

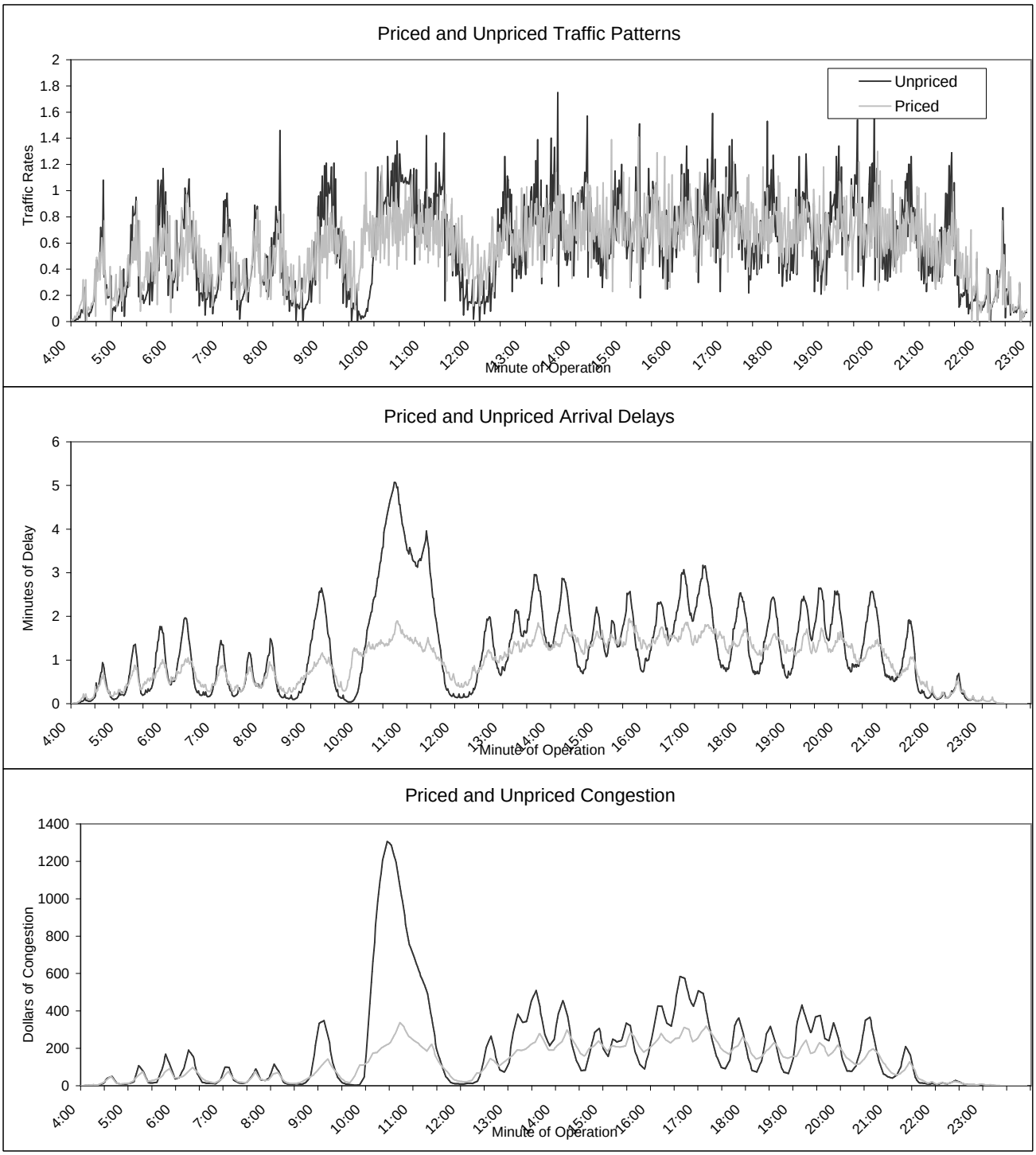

Figure 4--Non Internalizing Vancouver Model, Unpriced and Priced Arrivals, Delays, and Congestion

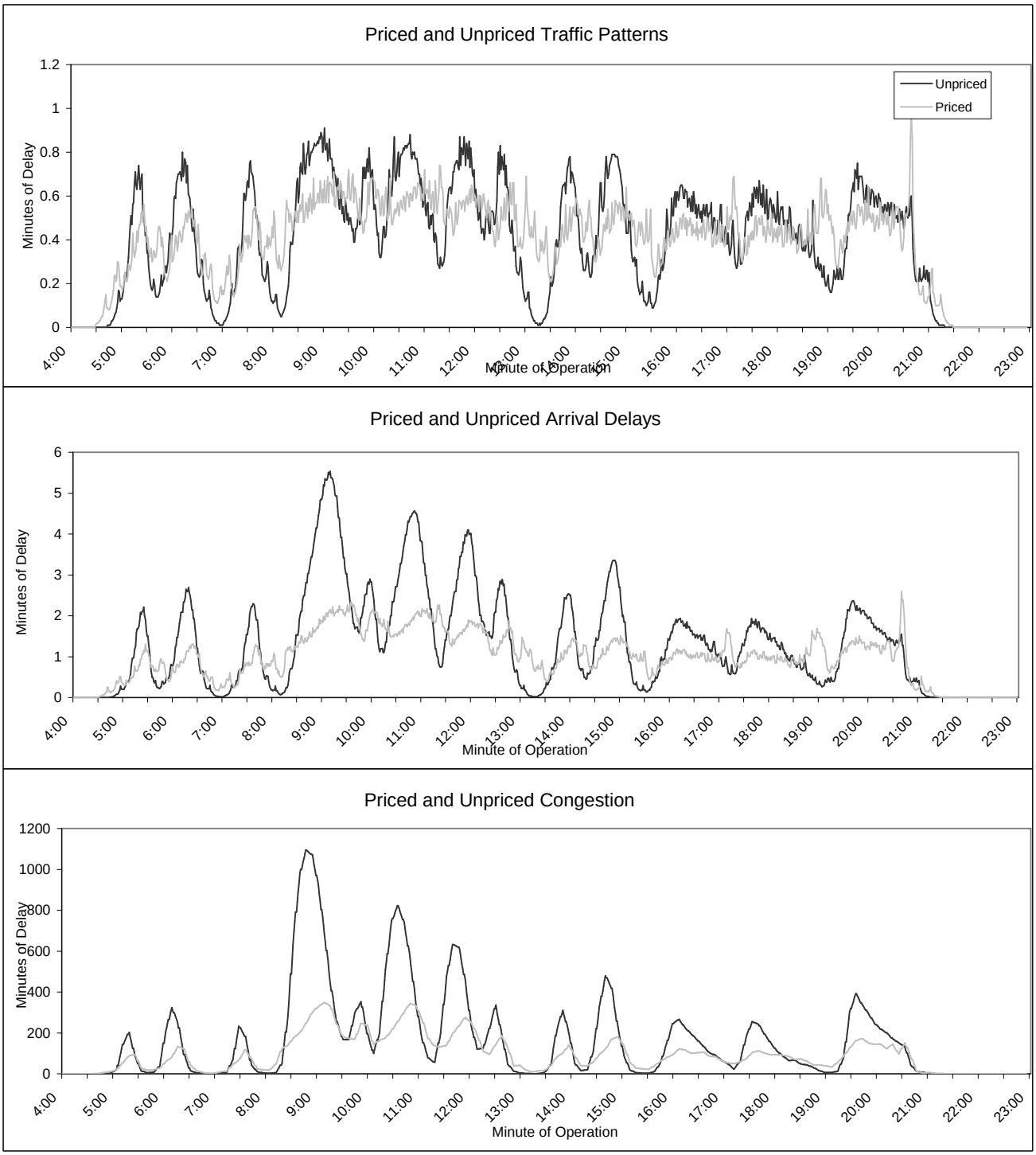

Figure 5--Non Internalizing Calgary Model, Unpriced and Priced Arrivals, Delays, and Congestion

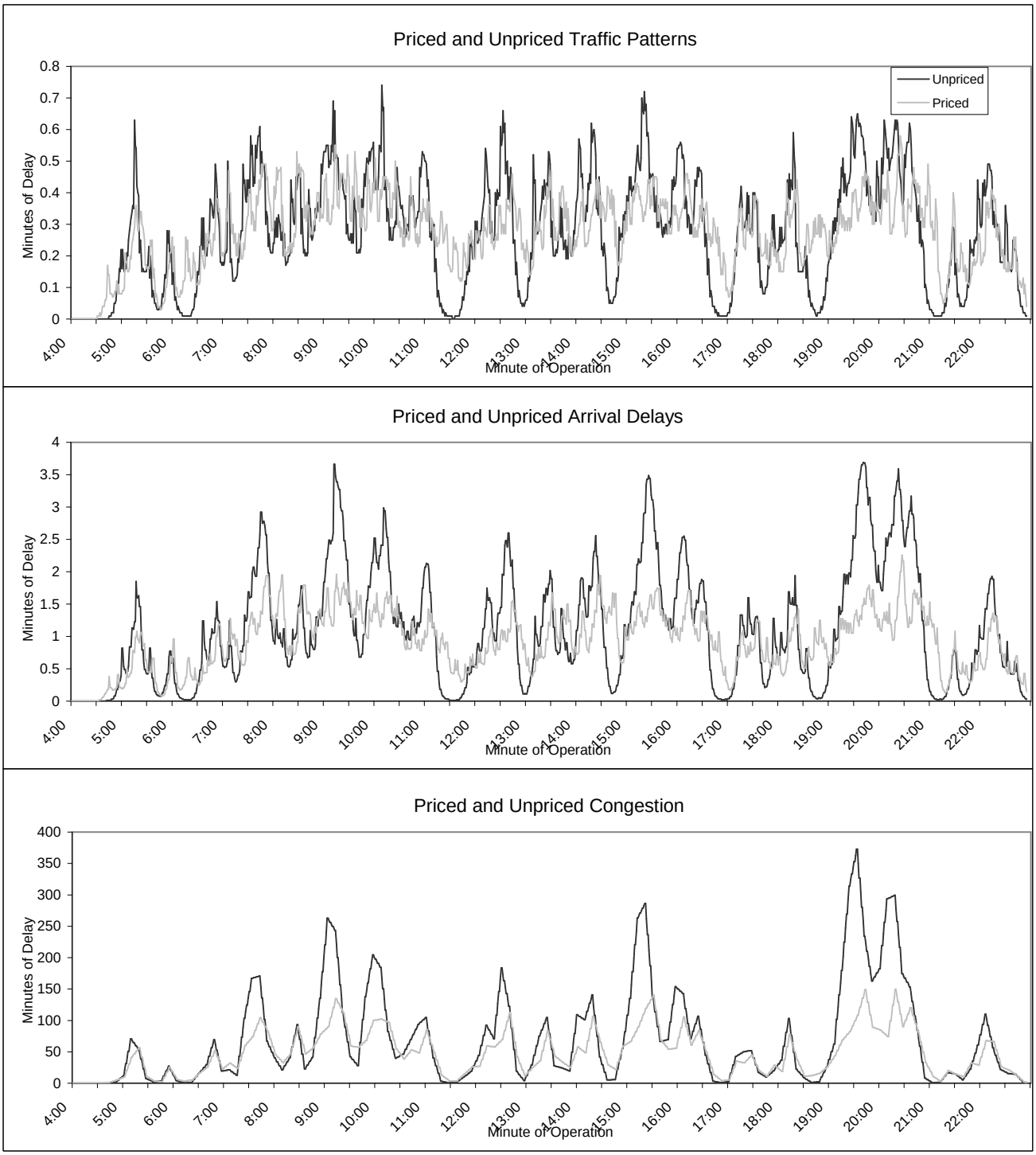

Figure 6--Non Internalizing Montreal Model, Unpriced and Priced Arrivals, Delays, and Congestion

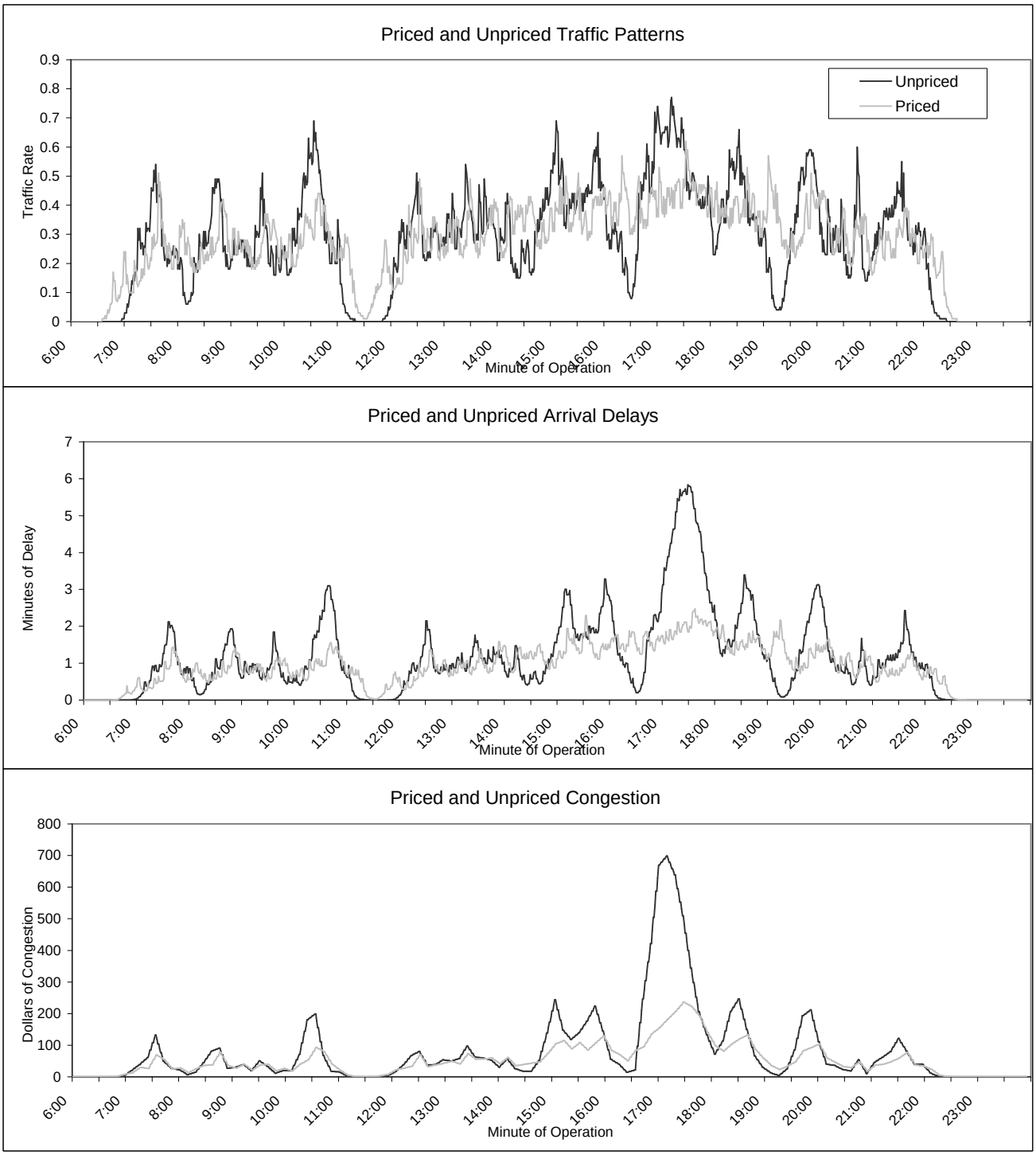

Figure 7--Internalizing Toronto Model, Unpriced and Priced Arrivals, Delays, and Congestion

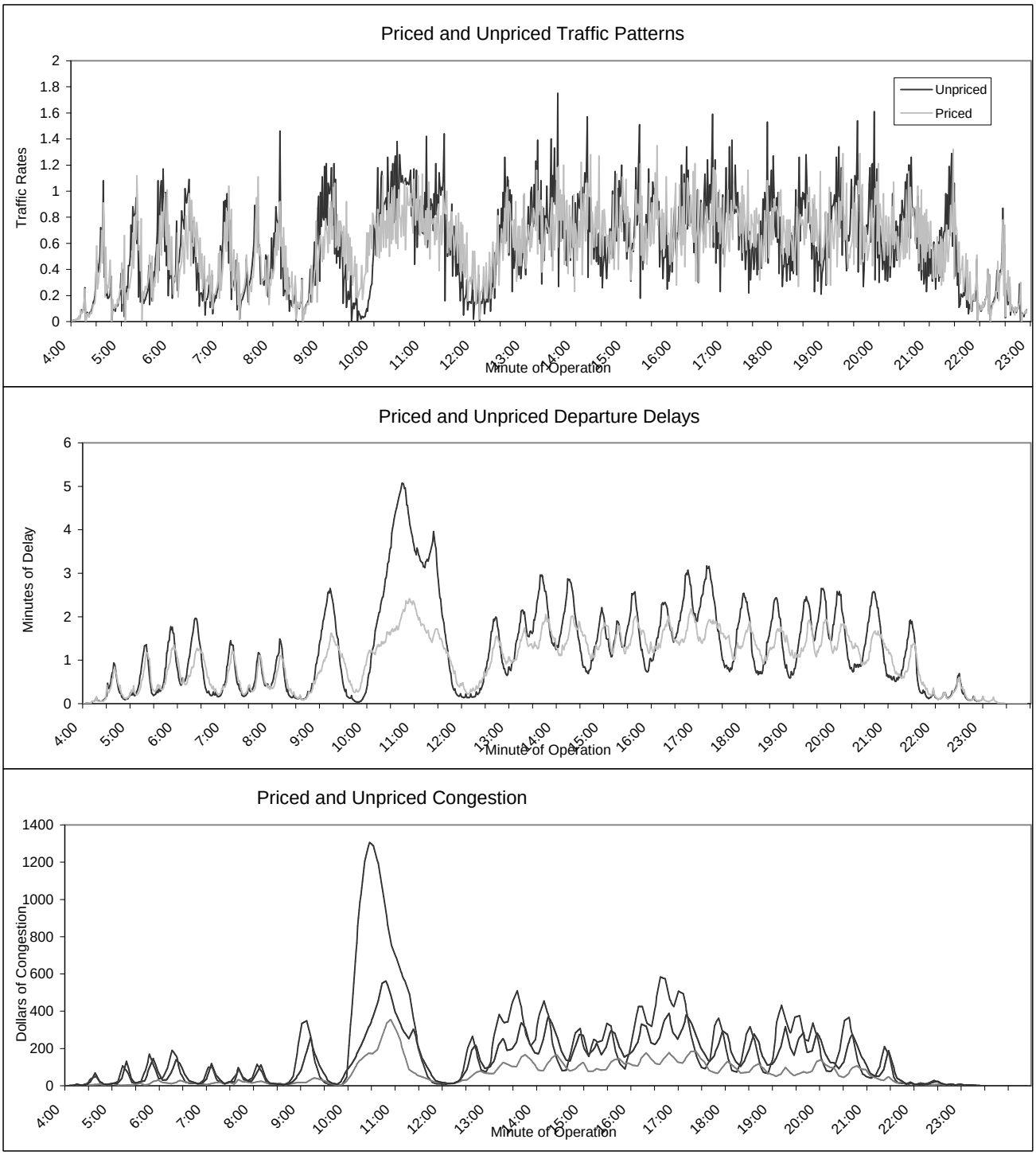

Figure 8--Internalizing Vancouver Model, Unpriced and Priced Arrivals, Delays, and Congestion

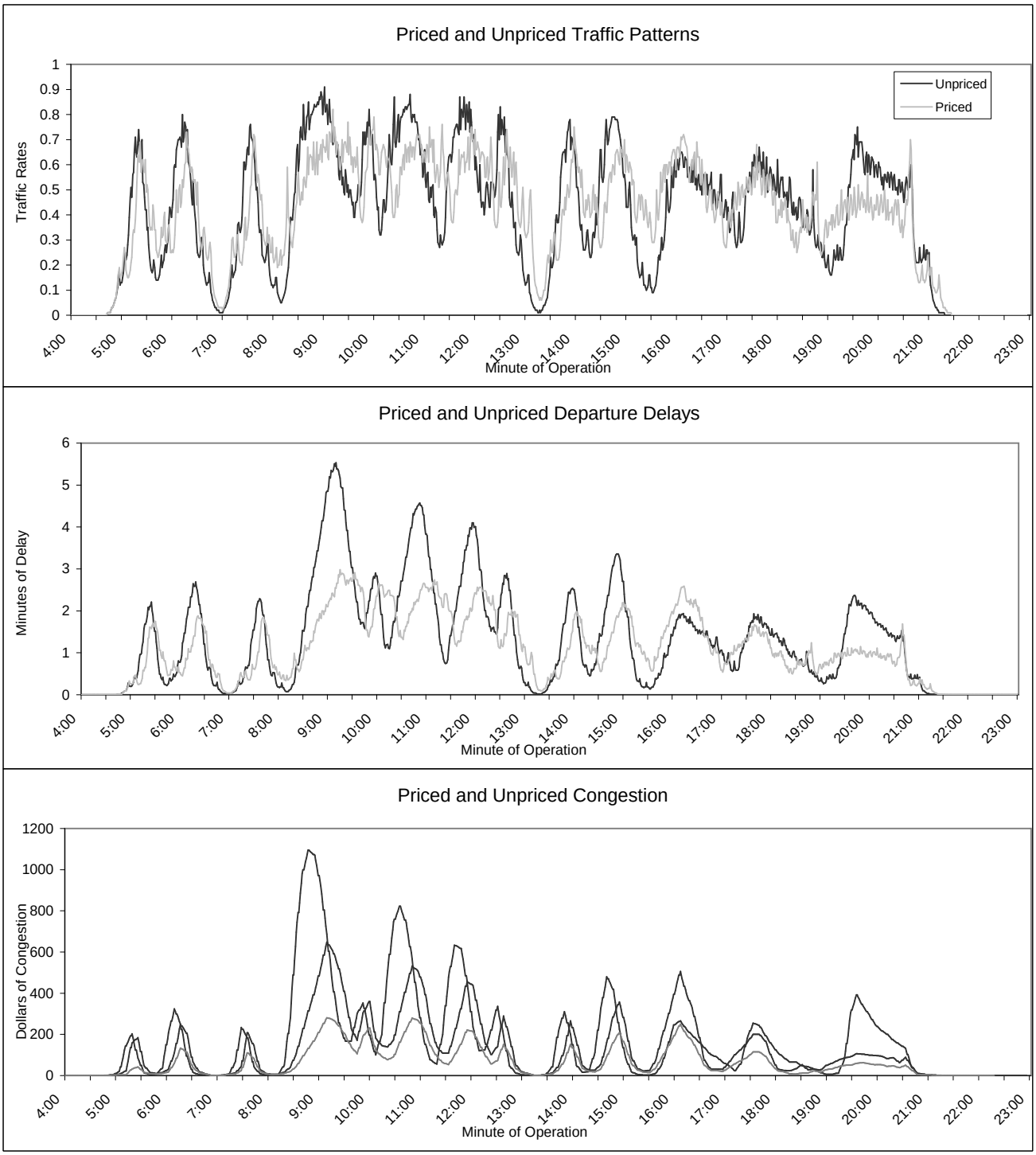

Figure 9--Internalizing Calgary Model, Unpriced and Priced Arrivals, Delays, and Congestion

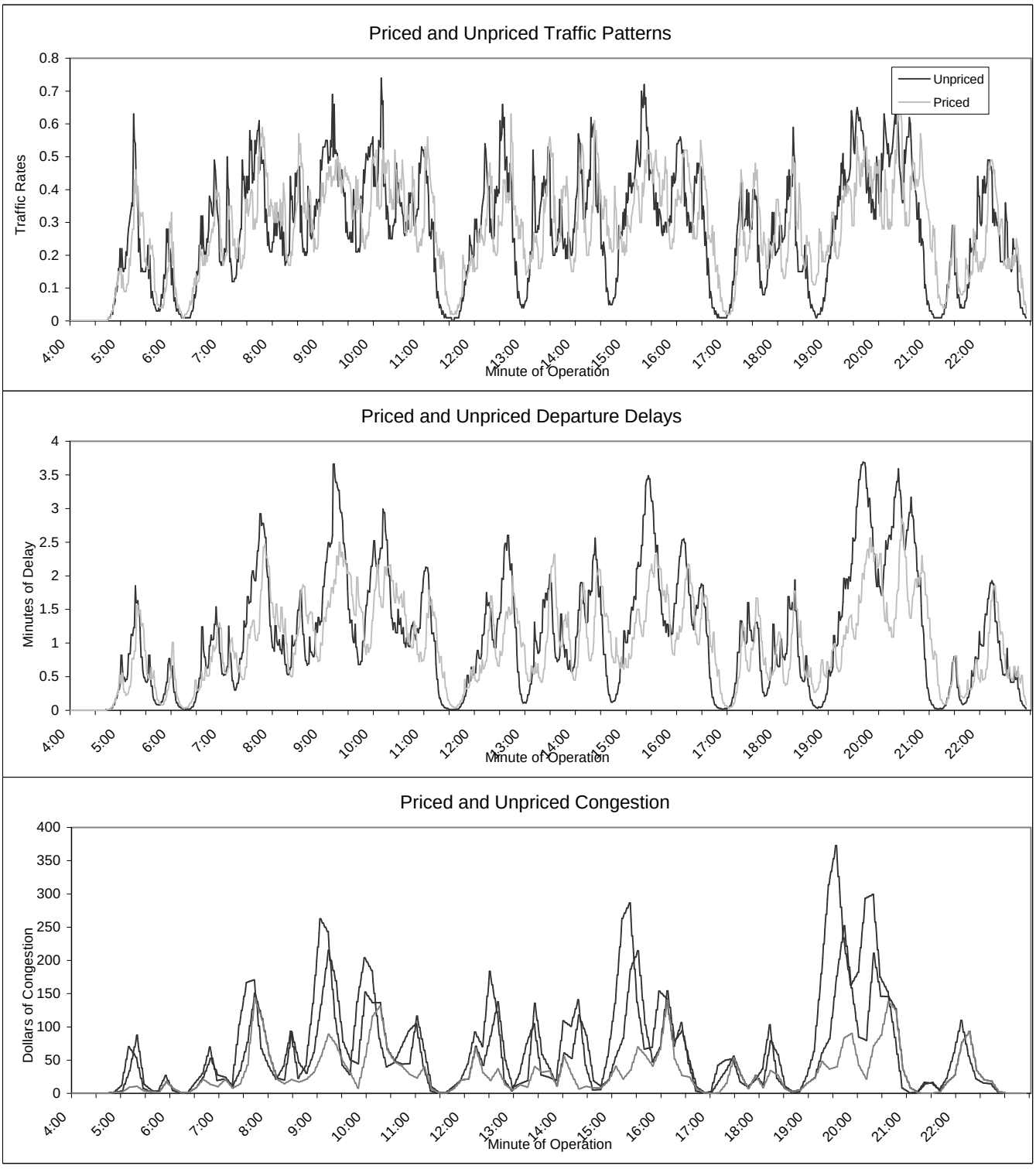

Figure 10--Internalizing Montreal Model, Unpriced and Priced Arrivals, Delays, and Congestion

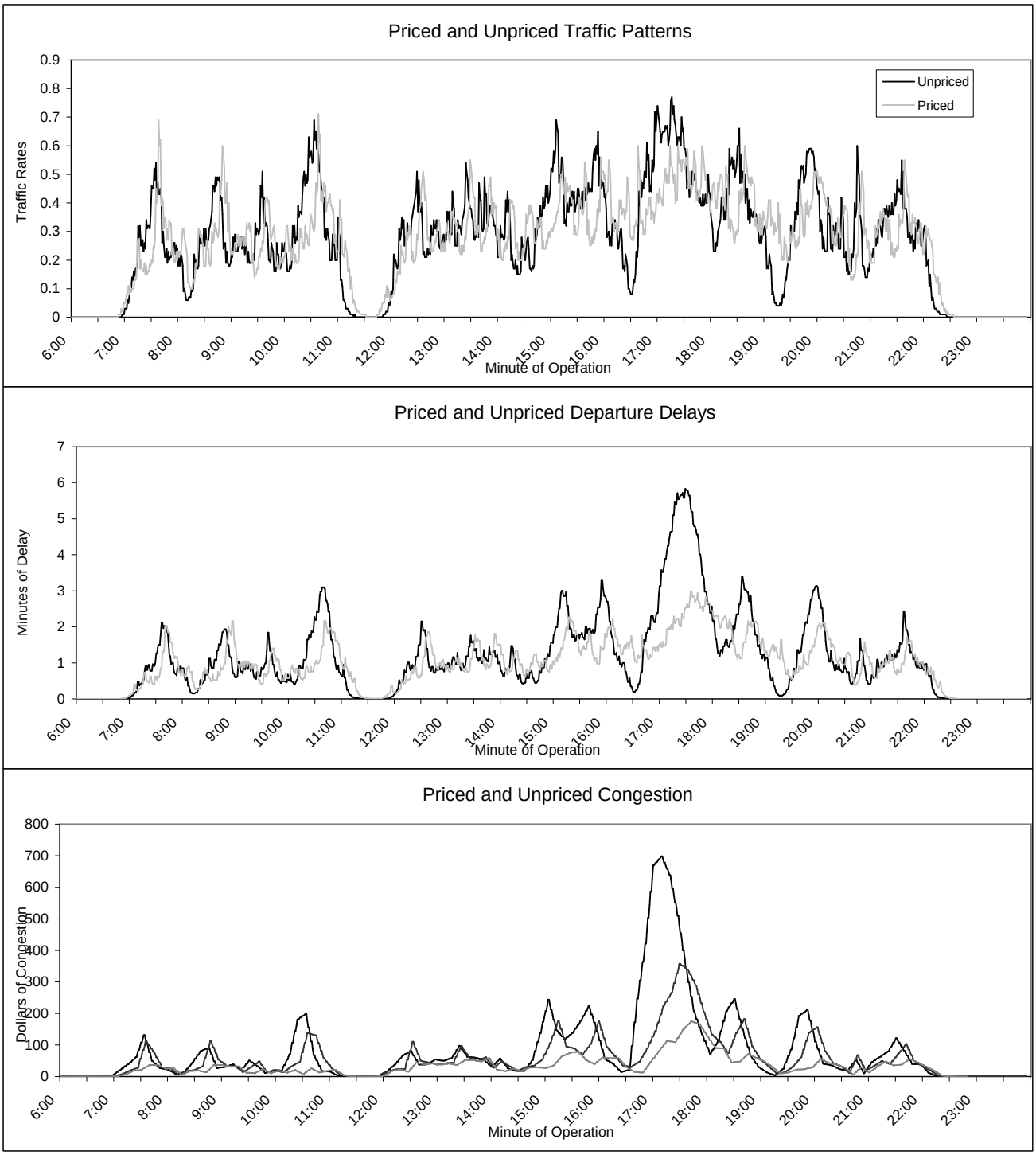

Figure A.1--Comparison of Scheduled Departures and Modelled Non Internalizing Departures

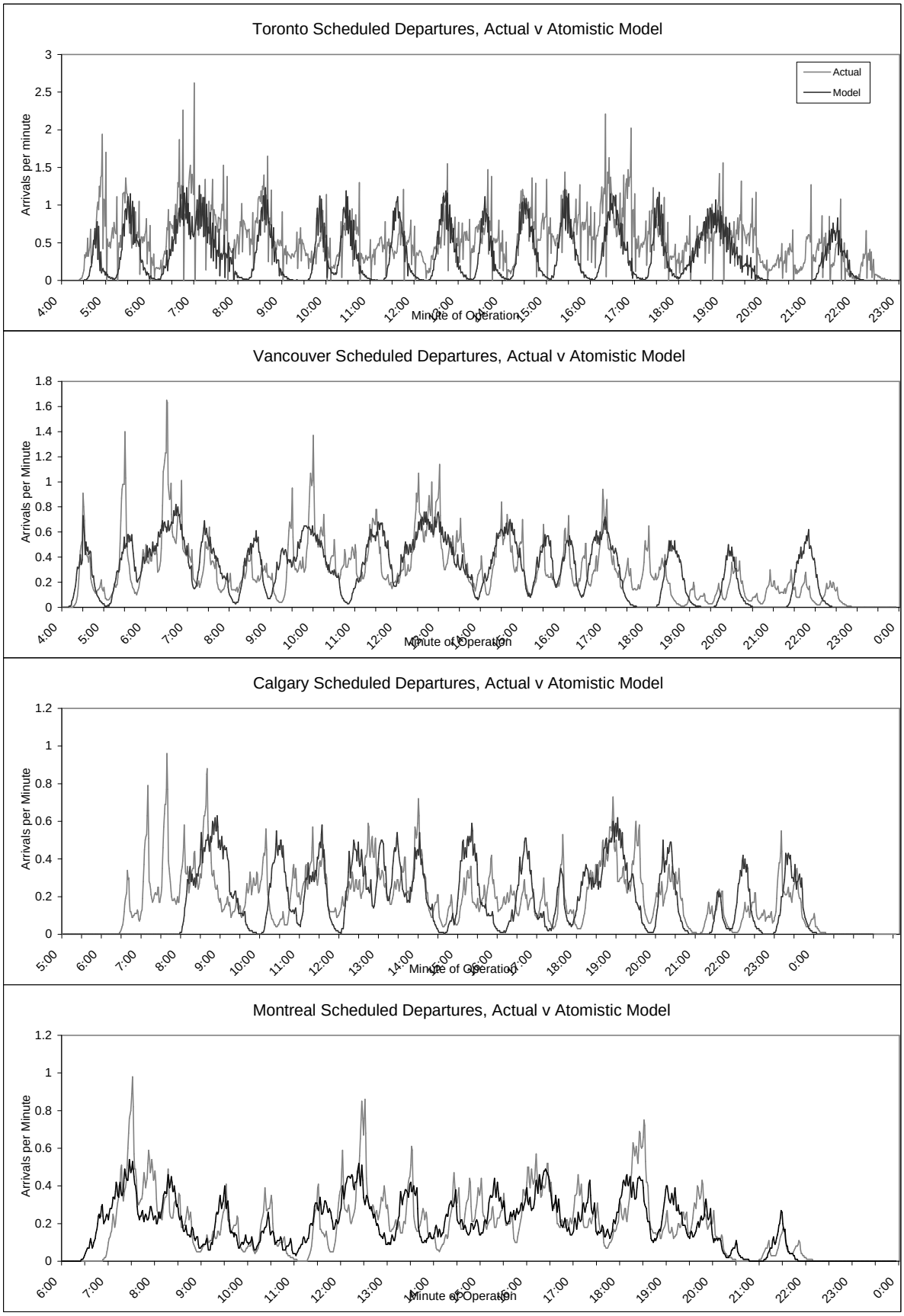

Figure A.2--Comparison of Scheduled Departures and Modelled Internalizing Departures